Modeling and mining of Web discussions

Anna Stavrianou

UNIVERSITE DE LYON

Ph.D. THESIS
of
Anna STAVRIANOU

prepared in the
LABORATOIRE ERIC - UNIVERSITE LUMIERE LYON 2

MODELING AND MINING OF WEB DISCUSSIONS

COMPOSITION DU JURY

M. Jean-Gabriel GANASCIA	Rapporteur	(Professeur, Université Paris VI)
M. Pascal PONCELET	Rapporteur	(Professeur, Ecole des Mines d'Alès)
M. Marc EL-BEZE	Examinateur	(Professeur, Université d'Avignon et des Pays de Vaucluse)
M. Stefan TRAUSAN-MATU	Examinateur	(Professeur, Université Politehnica de Bucarest en Roumanie)
M. Julien VELCIN	Co-directeur de thèse	(Maître de Conférences, Université Lyon 2)
M. Jean-Hugues CHAUCHAT	Directeur de thèse	(Professeur, Université Lyon 2)

Acknowledgements

Many people have helped in order to make the dream of this thesis come true.

First of all I would like to thank Prof. Nicolas Nicoloyannis who warmly welcomed me in the Laboratoire ERIC and accepted me as one of his PhD students. I will never forget his encouragement and kindness during the time we worked together. Unfortunately he left us suddenly in June 2007...

Since September 2007, the supervisor of this thesis has been Prof. Jean-Hugues Chauchat. His valuable ideas, comments, advice and guidance as well as the way he has managed this thesis have been extremely important and I sincerely thank him for this. I particularly thank him for accepting me as his student, for believing in me and supporting me at every moment during these years.

After September 2007, the co-supervisor of this thesis has been Dr. Julien Velcin. I would like to thank Julien for all what he contributed to this thesis. His enthusiasm, love for research, fruitful ideas have been significant. I thank him for the time he has spent re-reading papers, proposals and numerous other documents including this thesis as well as for insisting on certain important aspects of this work.

I would also like to acknowledge the scientific committee that evaluated my work, namely the professors Jean-Gabriel Ganascia, Pascal Poncelet, Marc El-Bèze and Stefan Trausan-Matu.

Working for the project "Conversession" gave me the opportunity to work with Mr. Robin Coulet whom I acknowledge for his ideas, feedback and precious comments especially on the system prototype.

All the people of the Laboratoire ERIC made these years enjoyable. Special thanks to Cecile for our fruitful discussions, Valerie for her help and professionalism and all my colleagues including Ahmad, Emna, Elie, Hadj, Hakim, Kamel, Marouane, Mathilde, Nora, Oksana, Remi and Sonia for helping me adapting to the lab life.

Special thanks to Emmanuel, François, Fred, Nikos, Markos and Maxime for helping out with the experiments. I really appreciate this!

Additionally I would like to thank Magda who advised me and supported me throughout the thesis. I also really enjoyed collaborating with Periklis in the beginning of the thesis. Thanks for all the tips!

No words can express how much I thank my husband, Fred. This thesis would not have been accomplished without his encouragement, understanding, support, love and patience. Big thanks for allowing me to follow my dreams.

I thank also my little boy, Giannis, who has been such a nice boy and has let his mum arrived to this point. I will never forget how nicely he surprised me when he saw a graph with red nodes in the prototype and told me with the few words he knew "ah maman, des fleurs...". He has undoubtedly changed my life. This thesis is dedicated to him.

Thanks to the crèche that welcomed my son and made me feel secure about my baby being there so that I continue my thesis.

Big thanks to my parents for their unlimited love and support, for everything they have done for me. I would have succeeded nothing without them. Also, thanks to my sister for always believing in me.

I would also like to thank my parents in law for their understanding and support all along these years.

I would finally like to thank all my friends for making me feel sure about my choices. You are all great!

Abstract

The development of Web 2.0 has resulted in the generation of a vast amount of online discussions. Mining and extracting quality knowledge from online discussions is significant for the industrial and marketing sector, as well as for e-commerce applications. Discussions of this kind encapsulate people's interests and beliefs and hence, there is a great interest in acquiring and developing online discussion analysis tools.

The objective of this thesis is to define a model which represents online discussions and facilitates their analysis. It has been partly implemented for satisfying the requirements of the project "Conversession" made for a start-up company supported by CREALYS. The objective of this company has been the management and analysis of online discussions.

In this thesis we propose a graph-oriented model. The vertices of the graph represent postings. Each posting encapsulates information such as the content of the message, the author who has written it, the opinion polarity of the message and the time that the message was posted. The edges among the postings point out a "reply-to" relation. In other words they show which posting replies to what as it is given by the structure of the online discussion.

The proposed model is accompanied by a number of measures which facilitate the discussion mining and the extraction of knowledge from it. Defined measures consist in measures that are underlined by the structure of the discussion and the way the postings are linked to each other. There are opinion-oriented measures which deal with the opinion evolution within a discussion. Time-oriented measures exploit the presence of the temporal dimension within a model, while topic-oriented measures can be used in order to measure the presence of topics within a discussion. The user's presence inside the online discussions can be exploited either by social network techniques or through the new model which encapsulates knowledge about the author of each posting.

The representation of an online discussion in the proposed way allows a user to "zoom" inside the discussion. A recommendation of messages is proposed to the user to enable a more efficient participation inside the discussion.

Additionally, a prototype system has been implemented which allows the user to mine online discussions by selecting a subset of postings and browse through them efficiently. Existing Text and Opinion Mining techniques have been integrated in the prototype system which demonstrates how the proposed model facilitates the mining of an online discussion.

Keywords: online discussions, opinion mining, text mining, social networks, recommender systems, modeling, forums.

Contents

List of Figures

List of Tables

Notations

G	a Post-Reply Opinion Graph (PROG)
V	the vertex set of a PROG graph
E	the edge set of a PROG graph
v, v', v_i	a vertex of the graph which represents a posting
$e_{v'v}$	an edge of the graph with direction from v' to v
m, m_v	a message in an online discussion
op_v	the opinion polarity encapsulated in the vertex v
u	a user who has participated in an online discussion
u_v	a user who has sent the posting represented by the vertex v
tm_v	the timestamp when the posting represented by the vertex v was posted
G_c	a discussion chain of the graph G
V_c	the vertex set of the discussion chain G_c
E_c	the edge set of the discussion chain G_c
G_{thr}	a discussion thread of the graph G
V_{thr}	the vertex set of the discussion thread G_{thr}
E_{thr}	the edge set of the discussion thread G_{thr}
T, T_i	a topic identified in an online discussion
n	negative opinion polarity
p	positive opinion polarity
o	neutral/objective opinion

Chapter 1

Introduction

The development of Web 2.0 has resulted in the generation of a vast amount of online discussions. The abundance and popularity of such discussions require an appropriate modeling for mining purposes. Mining and extracting quality knowledge from online discussions is significant for the industrial and marketing sector, as well as for e-commerce applications. Discussions of this kind encapsulate people's interests and beliefs and hence, there is a great interest in acquiring and developing online discussion analysis tools.

The objective of this thesis is to define a model which represents online discussions and facilitates their analysis. The thesis has been partly implemented for satisfying the requirements of the project "Conversession" made for a start-up company supported by CREALYS. The objective of this company has been the management and analysis of online discussions.

1.1 Motivation

For the purpose of understanding the motivation behind the needs of the analysis of online discussions, let us present three motivating examples.

1. A European politician has proposed a new law. This proposition has been discussed in plenty of online discussions, in various countries of Europe. The politician is interested in knowing what the reactions of the people are about this law. Navigating the discussions can give an idea of the people's reactions since in the web people express freely and anonymously their beliefs. Since it is impossible for the politician to

spend time on reading all postings in all existing discussions, a discussion analyst has been employed in order to fulfill the task of identifying what people say about this law. How will the analyst start accessing and analyzing the discussions? Which postings does s/he need to concentrate on and which ones to ignore? In the place of the politician and the new law, we could likewise have a company which desires to find out what people like and dislike about a new product.

2. A moderator whose job is to guide an online discussion wants to be able to know, at any chosen time, how the discussion has evolved. S/he wants to know, for example, which postings have caused a lot of reactions, in which parts of the discussion people argue, what are the main topics that appear in the discussion and whether the participants have shown interest in them. How can the moderator browse through the discussion and extract this information? How can we facilitate the comprehension of the online discussion?

3. An end user interested in the subject S has found an online discussion which discusses about this subject. The discussion is consisted of many postings which are split between many pages of the discussion site. The user desires to get quickly an idea of what the discussion discusses about and be able to participate as well. How will the user navigate through the online discussion quickly and efficiently?

The aforementioned examples are summarized in Figure 1.1. They present three scenarios in which people want to extract information from a discussion using various criteria such as topic and opinion knowledge. For the purpose of mining an online discussion such as a web forum and browsing through it, an appropriate representation is required.

Many current approaches [ARSX03, FSW06, JSFT07, STE07, ZAA07] represent an online discussion with a user-based graph, exploiting in this way the social network that is developed with the users of the discussion. As a result the information that can be extracted relates to how the users interact with each other. The semantic information as well as the structure of the discussion is lost.

When we deal with online discussions of the question-reply format such as a software forum where people search advice regarding how to implement certain tasks, then, the social network of users may reveal interesting information. One such information could be the identification of the experts

Figure 1.1: The desire of a user to mine online discussions.

[ZAA07] of the forum. The discussions, though, that provoke a conversation among users need to be represented in a different way which exploits the content and their structure. Therefore, we believe that a social network is not the most appropriate way for representing an online discussion for the following reasons:

- Inside an online discussion, when a message m_A is being replied to by a message m_B, the author of the reply message m_B intends to reply to the content of the message m_A and not to the author of that message. In other words, users reply to messages and not to users.

- In an online discussion the users do not know each other and they do not intend to get to know each other. They do not exchange messages in order to form a friendship or other kinds of relationships. They just want to express their opinions and have a discussion without noticing whom they are having the discussion with. What matters is the content of a message and not the identity of the user.

Therefore, an appropriate discussion representation is needed which facilitates the efficient browsing of a user inside an online discussion and the extraction of useful information such as:

- how many sub-topics appear in a discussion, what they consist of and whether some sub-topics tend to appear more than others,
- the postings which have caused many reactions or dispute over them, and which other postings may have influenced them,
- the average opinion of certain users and the opinion expressed against certain users inside the whole discussion or inside a particular sub-topic,
- whether the majority of reactions concern a negative or a positive position about a product/a law etc,
- the general evolution of the discussion.

The objective of this thesis is to provide answers to the presented scenarios and to propose a representation of online discussions which enables the aforementioned tasks. This representation should be structure and content-oriented and it should facilitate the browsing inside an online discussion, even if this is consisted of many messages.

1.2 Contributions

For the purpose of satisfying the requirements of extraction of information knowledge from an online discussion, we have developed a novel model. This model supports a graph-oriented representation where the vertices capture postings and the edges show the interaction among them. The process that is followed is depicted in Figure 1.2.

Initially the discussion we are interested in is being parsed and information about the postings is entered into a database.

The graph representation we propose consists of vertices which represent postings. Each posting encapsulates information such as the content of the message, the author who has written it, the opinion polarity of the message and the time that the message was posted. The edges among the postings point out a "reply-to" relation. In other words they show which posting replies to what as it is given by the structure of the online discussion.

The proposed model is accompanied by a number of measures which enable the browsing within the online discussion according to predefined criteria. Defined measures consist in measures that are underlined by the structure of the discussion and the way the postings are linked to each other.

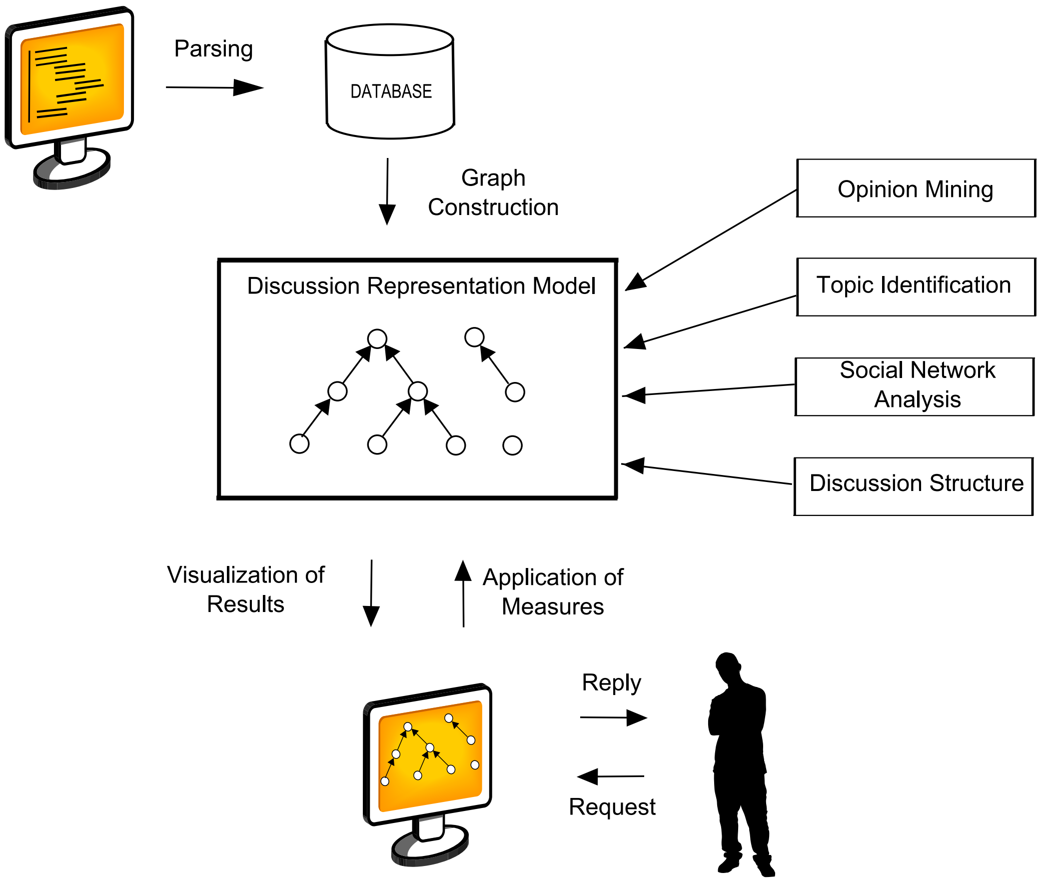


Figure 1.2: Analysis of an online discussion through our novel graph-oriented model.

There are opinion-oriented measures which deal with the opinion evolution within a discussion. Time-oriented measures exploit the presence of the temporal dimension within a model, while topic-oriented measures can be used in order to measure the presence of topics within a discussion. The user's presence inside the online discussions can be exploited either by social network techniques or through the new model which encapsulates knowledge about the author of each posting.

The representation of an online discussion in the proposed way allows a user to "zoom" inside the discussion. In this way information is provided regarding the positive or negative atmosphere of the discussion, the topics which appear, the participation of users and their opinion status, and the interaction between the postings. Moreover, a recommendation of messages is proposed to the user to enable a more efficient participation inside the discussion.

A prototype system has been implemented which allows the user to mine online discussions by selecting a subset of postings and browse through these postings efficiently. The prototype demonstrates how the proposed model facilitates the mining of an online discussion.

The contributions of this thesis are summarized in the following:

- **A novel model for analyzing online discussions.** We propose a framework for analyzing online discussions. The representation of a discussion is graph-oriented and it captures the structure of the discussion and semantic features of the content. Importance is given to the opinion presence inside the discussion. Currently most online discussion representations emphasize the presence of users and the interaction among them rather than the content or the structure itself.

- **Definition of measures.** We provide definitions of measures that are based on the proposed model. These measures use structural, opinion, temporal and topic information in order to facilitate the analysis of a discussion.

- **Recommendation of key messages.** We use the proposed graph-oriented model in order to define criteria which enable us to extract key messages from online discussions. Among the assumptions that lead to consider a message to be key are if the message initializes a thread, if it contains opinion or if it has caused many reactions. The criteria that satisfy these assumptions are correlated with what messages a user considers to be interesting. The extracted subset of key messages can be used in order to recommend messages to start-with to a new user inside an online discussion.

- **The System Prototype.** A prototype system has been implemented which allows the interaction of a user with an online discussion. The prototype enables the application of various criteria and it provides multilingual support when the appropriate Text and Opinion Mining plug-ins are available.

1.3 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2 discusses two very important fields that facilitate the analysis of online discussions; Text Mining and Opinion Mining. Both these fields deal with textual content and as such they are considered to offer a lot to the discussion analysis issue. In the scope of this thesis we use ready-made Text and Opinion Mining techniques in order to perform our experiments.

Chapter 3 presents our novel model for the representation and analysis of online discussions. Initially a brief overview is given of existing works which represent online discussions as social networks of users. Then, we proceed by formally defining our model, discussing its components and properties. We define measures which accompany the model and facilitate the discussion mining and the extraction of knowledge from it. Two examples are presented - an artificial and a real one - which demonstrate how the proposed model and its measures can be applied to online discussions.

Chapter 4 consists of an application of the model presented in this thesis. It regards the recommendation of key messages to the user in order to help him/her start having an idea of the discussion and find out whether there is interest in it and how to participate. The Chapter begins with a brief introduction of recommender systems and a discussion regarding the similarities between the recommender task and our approach. Some criteria based on the existing measures are proposed and their aggregation enables us to propose a subset of postings to the user to start with. Experiments with real users and forums in both English and French show that our model allows the definition of criteria which, when applied, can aid the recommendation process.

Chapter 5 presents the prototype which has been developed as part of this thesis. This prototype aids the user in viewing an online discussion as a graph, navigating through it, applying measures in order to better mine it and getting a recommendation of possibly interesting messages. The prototype is developed in such a way that is using plug-ins of existing Text and Opinion Mining methodologies.

Chapter 6 concludes this thesis and it presents various perspectives for future research.

Chapter 2

Text and Opinion Mining

Online discussions play a significant role mainly because of their content, since people use this kind of discussions in order to express their opinions and exchange ideas. In this thesis we have used existing Text and Opinion Mining techniques for the purpose of fulfilling the analysis task of online discussions. Text Mining refers to the discovery of previously unknown knowledge that can be found inside text documents. Opinion Mining is the field that deals with the presence of opinions inside a text such as the identification and extraction of opinions and arguments, the estimation of opinion polarities, the classification of a text according to its opinion tendencies. In this Chapter, we present an overview of existing approaches and methodologies regarding both fields. The Chapter begins by an introduction to Text Mining and continues in Section 2 by the motivation of the Text Mining field. Section 3 refers to Natural Language Processing issues. In Section 4 the focus is on the text representation techniques presented in the existing literature, while Section 5 deals with text categorization and the similarity measures used. Section 6 presents the Opinion Mining field while Section 7 discusses current methodologies that contribute to the identification of opinion polarities and tendencies inside text documents. Finally, Section 8 concludes.

2.1 Text Mining

The field of Text Mining has received a lot of attention due to the always increasing need for managing the information that resides in the vast amount of available text documents. Text documents, as opposed to information stored in database systems, are characterized by their unstructured nature. Sources of such unstructured information include the World Wide Web, governmental electronic repositories, biological databases, news articles, blog repositories, e-mails and, in general, any place where textual data is used for communication or reporting purposes.

Text Mining is the data analysis of text resources so that new, previously unknown knowledge is discovered [Hea99]. It is an interdisciplinary field that borrows techniques from the general field of Data Mining. Additionally, it combines methodologies from various other areas such as Information Extraction, Information Retrieval, Computational Linguistics, Categorization, Topic Tracking and Concept Linkage [FWRZ06, MB06].

It is often ambiguous to distinguish between the field of Information Retrieval (IR) and that of Text Mining. This happens because they both deal with text and its particularities, so they both have to face similar issues. IR has lent several algorithms and methods to Text Mining. The difference between these two fields is mainly their final goal. In IR, the objective is to retrieve documents that partially match a query and select from these documents some of the best matching ones [vR79]. Text Mining is about discovering unknown facts and hidden truth that may exist in the lexical, semantic or even statistical relations of text collections.

Another field that has lent methodologies to Text Mining is Information Extraction. Information Extraction differs from Text Mining because it regards the extraction of specific, structured data (e.g. names of people, cities, book titles) and pre-specified relationships [SAMK05] rather than the discovery of new relations and general patterns. In Text Mining the information found is unsuspected and unexpected, though in Information Extraction it is predefined and it matches the interest specified by the user [McC05, MB06, SAMK05]. Information Extraction techniques may be part of the Text Mining task in order to facilitate the knowledge extraction.

The Text Mining process consists of a data analysis of a corpus or corpora and it is concisely illustrated in Figure 2.1 [SAN07]. Taking a collection of text resources, a Text Mining tool would proceed with the data analysis. During this analysis many sub-processes could take place such as parsing, pattern recognition, syntactic and semantic analysis, clustering, tokenization and application of various other algorithms. Following the data analysis, the results are evaluated and the new, previous unknown knowledge may emerge. The retrieved text information can be used in various ways such as database population.

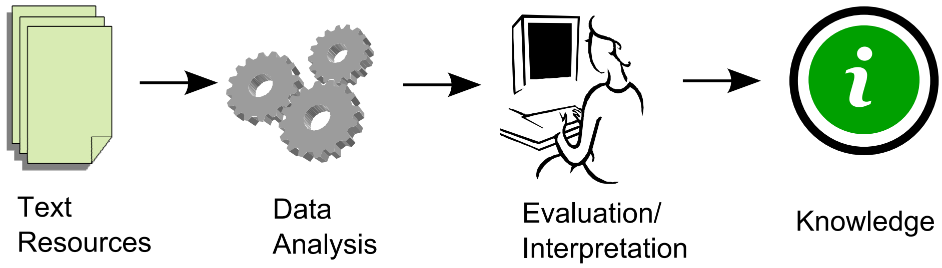


Figure 2.1: The Text Mining process.

A lot of Statistics and Machine Learning techniques exist and contribute to the data analysis, and therefore the Text Mining task. However, during the Text Mining process, many issues arise because of the automatic natural language processing (NLP) limitations, which the aforementioned techniques do not always take into consideration. A researcher needs to have a thorough overview of the existing difficulties posed by text before deciding on how to cope with them. In this Chapter we concentrate on the semantic issues present in Text Mining and we refer to some approaches that have attempted to handle these issues.

Throughout the Chapter, “terms”, “features” and “tokens” are used interchangeably according to context. The same stands for the words “text” and “document”.

2.2 Text Mining Motivation

The objective of Text Mining is the discovery of new knowledge within text collections. The magnitude of applications is significant.

In the biomedical field, most of the information is stored in text format so, association of terms and ideas is highly needed [ACK$^+$05, CH04, HPT$^+$02]. Swanson and Smalheiser [SS94, SS97] were among the first to observe linkages between text collections. This process led them to conclude a medical cause and effect hypothesis. The particular hypothesis was not then known in the medical academia, but it was later proved through scientific experiments. This shows that the analysis of correlations of information across text collections is advantageous in the biomedical sector. In this way unknown causes of diseases can be identified and, as a result, new medical treatments can be found. Of course, we should note that a lot of biomedical data is also stored in relational databases and the results of Text Mining can be used to facilitate further integration, update and querying of these sources.

Text Mining tools and methodologies have a lot to offer to data integration tasks. They enable the identification of similarities between text attributes that originate from different sources, reducing in this way the uncertainty and improving the data integration accuracy. Similarity measures in Text Mining extend beyond string-based similarity metrics. They may take into account syntactic and semantic information and they may be applied to words, phrases or even bigger pieces of text. The benefits of Text Mining to data integration during the merging of two companies can also be seen in [FWRZ06].

During data integration, issues such as record linkage and data cleaning are significant and they can also profit from the use of Text Mining approaches. Reducing redundant information and matching same entities across different sources and various representations, can be improved by using distance measures introduced in the Text Mining field. Semantics can help in dealing with incomplete information and erroneous data.

The applications of Text Mining can extend to any sector where text documents exist. For instance, history and sociology researchers can benefit from the discovery of repeated patterns and links between events, crime detection can profit by the identification of similarities between one crime and another [FWRZ06], and facts found in documents may be used in order to populate and update scientific databases.

Text Mining can definitely facilitate the work of researchers. It can allow them to find related research issues to the ones they are working on, retrieve references to past papers and articles which may have been forgotten and discover past methodologies that may add on the nowadays research. Text Mining may also reveal whether links exist between two different research domains without requiring the effort to understand the documents in both domains.

Another research field that may benefit from Text Mining is that of Information Retrieval since it is often required to execute queries that need the identification of semantic relations between texts. The application of Text Mining to Information Retrieval may also improve the precision of IR systems [Zai98] and reduce the number of documents that a single query returns.

Various other tasks can profit from Text Mining techniques. Examples consist of updating automatically a calendar by extracting data from e-mails [Fre98, McC05, WBMT99], identifying the original source of a news article [MBC+05], monitoring inconsistencies between databases and literature [NA06]. Finding out such inconsistencies requires the collaboration of database as well as Text Mining techniques. Missing database values could be filled in by data discovered and retrieved from the relevant literature.

2.3 Text Mining and Natural Language Processing

The particularities of the natural language pose many problems to Text Mining. In this Section, we will refer to the components of a language and the associated semantic issues.

A language consists of an alphabet, a grammar and a set of rules that define the syntax. The alphabet is the set of symbols used by a language. According to [Sha48], the letters and the sequences of letters have a statistical structure which means that they do not all appear with the same frequency. The grammar of a language is the set of rules that define how the symbols of the alphabet can interact with each other, while the syntax consists of the rules that capture the way the words can be united to form a sentence. According to Sapir [Sap21], "all grammars leak" since people tend to use the language freely, without adhering to rules. This stands, for example, for e-mails and online discussions, such as the ones dealt with in this thesis.

Describing text by a grammar can lead to erroneous identifications of lexical tokens, inability to capture syntactic text errors or identify certain items such as names [WBMT99]. Basic syntactic rules can though capture key patterns in the language structure. The syntactic rules depend on the language of the text and it is better if they are defined by linguists [MS99]. The rules may contain some uncertainty as in the case of a Probabilistic Context Free Grammar whose rules have probabilities attached to them.

Some of the natural language issues that should be considered during the Text Mining process are listed in Table 2.1 and they are discussed in this Chapter.

Stop List. Using a stop list which usually contains high frequency words such as 'a', 'the' or 'of' that are to be ignored from a text, is an idea inherited by the Information Retrieval field, where it has been widely used due to improved retrieval results. In Text Mining, though, it is not as useful since

Table 2.1: Text Mining issues that need to be considered by researchers before they proceed with the mining of a text.

Issue	Details
Stop List	Should we remove or take into account stop words?
Lemmatization / Stemming	Should we reduce the words to their lemmas or let them as they are?
Noisy data	Should the text be clear of noisy data such as orthographic mistakes and abbreviations?
Word Sense Disambiguation	Should we clarify the senses of words in a text?
Tagging	What about data annotation and/or part of speech characteristics?
Collocations	What about compound or technical terms?
Grammar / Syntax	Should we make a syntactic or grammatical analysis? What about data dependency, anaphoric problems or scope ambiguity?
Tokenization	Should we tokenize by words or phrases and if so, how?
Text Representation	Which terms are important? Words or phrases? Nouns or adjectives? Which text model should we use? What about word order, context, and background knowledge?
Automated Learning	Should we use categorization? Which similarity measures should be applied?

common terms seem to provide information [Ril95, YP97]. Common stop words can even help in clarifying the semantics of a text segment.

Lemmatization. Lemmatization is dependent on the language of the text. It reduces a word to its root e.g. it replaces "reading" or "reader" by "read", so that similarity detection can be achieved. Applying lemmatization or stemming techniques to a piece of text may affect the semantics.

Noisy data. Correcting spelling mistakes and replacing acronyms and abbreviations can also be part of the Text Mining process in order to eliminate noisy data before the main processing starts. During this text cleaning, the use of a dictionary or thesaurus may be useful. The text cleaning, here, differs from the data cleansing in the databases field in that it is mainly about misspellings rather than schema inconsistencies, integrity constraints or invalid data.

Word Sense Disambiguation. The word sense disambiguation (WSD) issue is about finding out the most probable meaning of a polysemous word. One approach to solve this is by considering the context in which a particular word is found. This process may include obtaining the grammatical category of a word, for instance, detecting if the word "play" is a noun or a verb in a specific phrase. There are two types of disambiguation; the supervised and the unsupervised. The supervised one is often carried out with the help of a dictionary or a thesaurus. In the unsupervised disambiguation, the different senses of the word are not known. Yarowsky [Yar95] has presented an unsupervised approach to the WSD problem with high accuracy results.

Tagging. Tagging concerns the application of part of speech (PoS) tags, XML or SGML mark-up to corpora. PoS tags capture certain syntactic categories such as nouns, verbs and adjectives, and they can be used for the identification of noun phrases or other parts of speech. In case unknown words exist in a text, there are ways to find the most probable tags since the possibility of some tags having unknown words is not the same for all of them [MS99]. The Brown corpus (http://helmer.aksis.uib.no/icame/brown/bcm.html) and the Penn Treebank (http://www.cis.upenn.edu/ treebank/home.html) are well -known text collections that are tagged by grammatical tags.

Collocations. Another issue is that of the collocations that may exist in a text. These are phrases, such as "radio therapy", that make sense only if considered as a whole. In collocations, the meaning of the whole is greater than the meaning of the sum of its parts. In other words, the semantics of a collocation are not equal to the semantics of its parts, so studying the properties of the single words does not convey the meaning of the collocation itself. A syntactic analysis may lead to collocation discovery in a text.

Syntactic Analysis. If a syntactic analysis takes place, the order in which the words appear in the text is an issue that should be considered. The parsing of a sentence could start either by the beginning or by the end of it and sometimes it could even start by the main verb since this usually directs the development of a sentence.

Tokenization. Tokenization regards the splitting of a text into units and it may take place during the data analysis. A text can be tokenized in paragraphs, sentences, phrases of any length and single words. The delimiters used vary. A common delimiter is the space or the tab between words. Punctuation marks can be used as well, such as full stops, exclamation marks or commas. Particularities of the delimiters may need to be considered. For example, the full stop is used in abbreviations so apparently it does not always mark a sentence ending. Also, considering the space as a tokenization symbol will keep the compound phrases apart.

Common stop words such as "and", "the" or "a" can be considered as delimiters [BP01] or even specific domain stop words (e.g. technical terms) dependent on the domain the text belongs to. The terminology is a sensitive issue whose extraction has been dealt with in some papers [Bou92, DGL94]. Bourigault [Bou92] defines the technical terms as noun phrases which have a meaning even if they exist outside a text.

Tokenization can also be done in paragraphs or sections. This is often referred to as discourse segmentation. In [Koz93] text segments are found by calculating the lexical cohesion between word lists. Changes in the lexical cohesion can be considered as segment boundaries. Another example is the TextTiling algorithm [Hea94] which partitions a text into subtopics. The algorithm splits the text in phrases of certain length, it checks the term repetition and the lexical similarity between these phrases, and it defines the thematic boundaries wherever the similarities change dramatically. The evaluation of this algorithm shows that human judgment is reflected in the way the segmentation is done.

2.4 Text Representation

Similarly to database models, text models intend to capture the relationships between data. Text models, though, describe free text and not structured data. The relationships may be derived by statistical ways and not necessarily through logical associations. Moreover, the operations of a text model are usually between vectors and the data do not comply with a logical schema.

Text representation may serve as an intermediate step between raw text data and database models. For example, organizing data found in documents into relational tables requires some text and semantic analysis that is applied on text models. Database models are used for data storage and curation, while text models permit the discovery of similarities among texts, topic identification and text linkages that may not be obvious.

The most widely used representation is the Vector Space Model (VSM) [SWY75]. According to this, the text is described by a vector whose dimension is the number of text features and its content consists of a function of the frequencies with which these features appear in the corpus or corpora. This model is also referred to as the bag-of-words model because the order and the relations between the words are ignored.

The majority of representations proposed are an extension of the VSM model. There are some representations that focus on phrases instead of single words [BP01, CMS01, MBSC97], some that give importance to the semantics of words or the relations between them [CHS05, KPKF01, RB99] and others that take advantage of the hierarchical structure of the text [AG06]. These different approaches are discussed in the following sections.

2.4.1 Feature Extraction

A lot of discussion dating back to the Information Retrieval field concerns whether frequent or rare terms are more suitable to represent a text and whether single words or phrases are better terms.

The frequency with which a term appears in a corpus or corpora can clarify the significance of this term in a specific document. A frequency measure can be binary to underline absence or presence, it can vary from 0 to 1 or it can be given by a mathematical function. Normalization is usually needed so that the length of the document and the number of unique terms is taken into account. For instance, in a very small text that contains only 10 unique terms, all the terms are important regardless of their frequency.

An example of a statistical index that gives a quantitative answer as to whether a term, being frequent in one document, is really worth being extracted when it is also frequent in a collection of documents is the well-known *tf-idf* index. This index promotes terms that appear many times in a single document but very few times in a collection of them [Seb02].

Statistical information can be gathered either for distinct words or phrases. Lewis [Lew92] supports that words provide better statistical quality. This is because the words which constitute a phrase may appear multiple times in a document while the phrase itself may be present only once and as a result the frequencies can be misleading.

On the other hand, phrases provide more semantic information than the single words because they give an idea of the context. A word is characterized by the company it keeps [Fir57] and since words may have multiple meanings, we do need to know at least the phrase that contains the word in question, so as to approach the semantics with higher certainty. The experiments of Blake and Pratt [BP01] demonstrate the benefit of using special phrases and concepts over words for the representation of medical texts.

The interest in collecting statistical and semantic information has led to the issue of choosing between statistical and syntactic phrases [CMS01, MBSC97, Seb02]. A statistical phrase is a phrase that appears in a statistical way inside a text, while a syntactic one is a phrase whose grammar and syntax rules reveal some semantics. A statistical phrase is retrieved by statistical methods while a syntactic phrase can be extracted using linguistic methods.

Salton [Sal88] combines statistical and syntactic phrases for book indexing. He carries out a syntactic analysis of the sentences of a document and then he extracts from the syntactic tree some of the existing noun phrases. He gives importance to the frequency of terms within a document and within a collection of documents and he marks the noun phrases of the document title.

In Table 2.2, the advantages and disadvantages of considering words or phrases as terms are shown.

When we have to make a decision between using words or phrases, the important is not which kind of phrases is better but whether they have to offer something more than the single terms [FMR98, MBSC97]. As it can

Table 2.2: Advantages and disadvantages of using words and phrases as features.

	ADVANTAGES	DISADVANTAGES
WORDS	good statistics, synonyms, existence of tools or algorithms (e.g. WordNet, WSD algorithms)	no context information, problem with collocations
PHRASES	context information, semantic quality, collocations can be captured	average statistical quality

be seen from Table 2.2, phrases fill in the gaps that words cannot cover and vice versa. Phrases inform about the context, while words provide higher statistical quality. Therefore, it seems that a combination of both is the best way to capture text features.

2.4.2 Representation Models

The VSM model can only capture information related to the frequencies of text features. Alternative models have been proposed in the existing literature using a variety of features such as words, phrases and concepts and applying data such as the knowledge of word senses. The representation models vary between the use of n-grams, vectors, trees and other type of hierarchies.

Context of features. The context of a term is a useful piece of semantic information. Rajman and Besançon [RB99] have represented the context as a vector that contains the co-occurrence frequencies between a term and a predefined set of indexing features. Nenadic and Ananiadou [NA06] use context patterns in biomedical documents. These patterns are in the form of regular expressions and they contain PoS tags and ontology information.

N-grams can also be used to discover the context of a word. Caropreso et al. [CMS01] have used n-grams in order to represent and categorize text. They replace some unigrams with bigrams and they use functions such as document frequency and information gain in order to score the n-grams extracted from the text. Their results are better when bigrams are used over unigrams. Similar results have been shown in [MG98]. N-grams can refer to either a sequence of n words or n letters. The second case is well adapted for multilingual texts which contain words in more than one languages [JCD04].

Cimiano et al. [CHS05] model the context of a term as a vector of syntactic dependencies found in a text corpus. They extract a concept hierarchy by applying a method based on the formal concept analysis. A linguistic parser extracts the syntactic dependencies. Then, they assign weights to these dependencies and they create a lattice of formal concepts. The problem is that the size of this lattice increases according to the number of concepts.

Sense of features. Kehagias et al. [KPKF01] have experimented by using sense-based representations where the features chosen are not single words but the meanings of them. The results of the research have not shown improvement in the accuracy of text classification compared to the accuracy achieved by the word-based representations.

Hierarchy. Carenini et al. [CNZ05] propose a hierarchy of extracted features. They attempt to map texts that describe product reviews to a UDF (user-defined features) hierarchy. The advantage of using such a taxonomy, as it is reported in the paper, is adding background user knowledge to the model and reducing the redundancy. The disadvantage is that for every (sub-) domain a UDF hierarchy has to be created.

Similarly to [CNZ05], Bloehdorn et al. [BCH05, BH04] match the syntax of sentences found in a text against a library that contains regular expressions patterns. The concepts found are added to the bag-of-words model creating in this way a "hybrid feature vector".

Matrix space models (MSM) have been proposed for text representation [AG06]. This representation is based on the idea that a document is a hierarchy of document extracts e.g. sections, paragraphs and sentences and as a result term-by-section, term-by-paragraph and term-by-sentence matrices can be respectively created. In [AG06] they deal with term-by-sentence matrices. Their experiments regard query evaluation for IR and the results are close to the ones achieved by Latent Semantic Indexing (LSI) with low computational cost. Accuracy is said to be high for multi-topic documents. The advantage of this kind of matrix representation over the VSM and the LSI model is that it "remembers" the intermediate steps of the construction of the final matrix.

A structured text having sections, paragraphs and sentences is better than a totally unstructured set of words [Koz93]. Therefore, considering text properties such as the location of a word in a text can lead to a better representation. The words present in the title of a document have usually higher significance. It can also be considered that the first paragraph of a document is often an introduction while the last one is usually a conclusion.

In Table 2.3, we present some of the approaches covered by the existing literature together with the text units they focus on, the representation types they use and the task they are dealing with.

Table 2.3: Text Representation Approaches.

Approach	Terms	Representation Type	Objective
Antonellis and Gallopoulos [AG06]	Sentences	Term-by-sentence matrices	Text Mining
Blake and Pratt [BP01]	Words, phrases, concepts	Association rules	Representation of medical texts
Bloehdorn et al. [BCH05, BH04]	Words and concepts	Combination of bag-of-words and concept hierarchy	Text clustering and classification
Carenini et al. [CNZ05]	Concepts	Hierarchy	Feature extraction
Caropreso et al. [CMS01]	Phrases	N-grams	Text categorization
Cimiano et al. [CHS05]	Concepts	Concept hierarchy	Automatic acquisition of a taxonomy
Kehagias et al. [KPKF01]	Word senses	Sense-based vector	Text categorization
Mladenic and Grobelnik [MG98]	Phrases	N-grams	Text learning
Rajman and Besanon [RB99]	Words and compounds	Vector	Information Retrieval
Salton [Sal88]	Noun phrases	Tree	Book indexing
Salton et al. [SWY75] (VSM)	Words	Vector	Information Retrieval

2.5 Categorization

The data analysis of corpora often involves the identification of the inherent structure of the document collection, the labeling of documents and text segments and the generation of clusters according to a similarity measure. The task that deals with the organization of an unstructured collection of documents to a structured repository is called text categorization and it aims at facilitating storage, search and browsing [Seb06].

The categorization task can be supervised or unsupervised, dependent on whether the groups or categories are known from the beginning or not.

During a supervised classification process, the first step is to define the documents that will be used. There are three sets of documents; the training set with annotated documents, the development set used to test the classifier before it is completed, and finally, the test set that comprises the documents which will evaluate the performance of the classifier. The intersection of these three sets should be the empty set. Subsequently, the representation of these documents and categories is decided. The training of the model begins, the parameters are tuned and the model is applied to the test documents. The computational cost of text annotation and the difficulty in obtaining training data, has led the researchers to alternatives such as semi-supervised techniques [AZ05, CLWL04, NG00] that use a small set of labeled data.

In the unsupervised case which is called clustering, there are no labeled documents. A similarity measure is defined and the documents are compared with each other in order to be divided into clusters. The objective is to achieve a low inter-cluster and a high intra-cluster similarity.

The text categorization algorithms can be applied in many cases. The thematic labeling of a document collection, the classification of movie text reviews into positive and negative ones, the distinction of spam e-mails from the rest and the automatic organization of Web pages are examples of categorization. In this Section, the word "categorization" is used to refer to both supervised and unsupervised cases.

2.5.1 Categorization Tasks

The categorization task may vary according to the intra-document or inter-document associations that need to be captured. Thus, the categorization goal should be clear before deciding which algorithm to apply. The goal can be the identification of the documents that deal with the same topic, the semantic orientation of a review, the selection of the articles written by the same author, the disambiguation of the meaning of a polysemous word in a text or even the distinction between interesting and not interesting texts based on the preferences of a person. In the existing literature, various categorization cases have been considered. Here we briefly discuss some of them.

Topic categorization. In the case of topic categorization, i.e. a classification of documents according to their topic, the focus is usually on noun terms that may characterize a topic. The techniques to identify the topic of a text vary from simple ones such as considering as topics the words that appear more frequently in the text to more advanced ones such as machine learning methods.

Trausan-Matu and Rebedea [TMR09] have performed topic identification in online discussions that are in the form of chats. Apart from using the frequency of words as a method for identification, they observe that there are standard expressions that point out the initiation of a new topic such as "let's talk about..." or "what about...". These expressions can help in finding out when a topic is initialized.

Regarding the machine-learning methods, Sebastiani [Seb02] takes into account the experiments presented in various articles. His conclusion is that boosting-based [Sch99], example-based (e.g. k-NN), based on regression methods (e.g. LLSF) classifiers and SVM are regarded as top classifiers. Neural networks and online linear classifiers (e.g. perceptron) follow the aforementioned top ones and they are considered to be very good. Recently, the Latent Dirichlet Allocation [BNJ03] model has been proposed in order to point out which topics are discussed in a document collection.

In the field of topic categorization, we also have the approach AGAPE which focuses on the extraction of concepts from a set of textual data [VG07]. AGAPE is an unsupervised approach and it uses no knowledge about the existing number of clusters. It assigns one single topic which is defined as a set of words or phrases to each textual input. In addition, each word or phrase can belong to only one topic.

Sentiment categorization. A sentiment classification task deals with the classification of a document according to the subjective opinion of the author [JB06]. In this case, the focus is on finding the semantic orientation of a word, namely its positive or negative attitude. Sentiment identification techniques are discussed later, in the Opinion Mining Section.

Sentiment classification seems to be more difficult than the topic-based one and it cannot be based on just observing the presence of single words. In [TL03] it is mentioned that sarcasm may be an obstacle for the clarification of the semantic orientation of a text. More sophisticated methods need to be employed so as to differentiate between the subjective and objective opinion of a reviewer or between the objective description of a movie and references to other people's comments. An initial step in recognizing subjective and objective statements is presented in [JB06] where they focus on identifying comparative sentences.

2.5.2 Measuring Similarity

Another part of a categorization task is the selection of a similarity measure in order to identify the mutual characteristics of various documents. Dissimilarity measures, which focus on how dissimilar two concepts are, may also exist. Any dissimilarity function can be transformed into a similarity one but the opposite does not always stand [vR79]. The similarity measures proposed in the existing literature can be divided into two categories; the statistical and the semantic ones.

Statistical Measures. Measuring the term frequency and the co-occurrence frequency has been widely used. According to Resnik [Res95], the co-occurrence frequency is a proof of relatedness. Hoskinson [Hos05] uses a combination of document co-occurrence and term frequency measures in order to classify concepts which are defined as the most frequent terms. Among the most popular statistical measures are the cosine coefficient, the Euclidean distance and the chi-square which are used by text classifiers in order to compare two vectors.

Semantic Measures. The semantic-based similarity measures the distance between the meanings of two terms. WordNet (http://wordnet.princeton.edu/) is often used in order to find out word senses or semantic relations between wording features. It is an electronic database of the English language that consists of words organized into subsets according to their meaning. These subsets are synonym sets called synsets, and they are linked by relations such as inheritance or part-whole relationships.

For languages other than English, there are some projects found in the Global WordNet Association web site (http://www.globalwordnet.org/) such as EuroWordNet (http://www.illc.uva.nl/EuroWordNet/).

Measuring the similarity between two nodes in WordNet or a similar hierarchy can be done in many ways. The edge-counting method measures the path length from one node to another. To avoid problems that appear by not taking into account the density of the hierarchy, an information content measure has been used [Res99, SVH04] in some cases, showing improvement in the results. The information content measures the amount of information that can be given by a concept or a term. The more abstract a concept is in a hierarchy, the higher it is and the less information it contains. As a result its information content has a low value. Additionally, the more information is shared between two words or two concepts, the more similar they are.

Budanitsky and Hirst [BH01] have compared some similarity WordNet-based measures concluding again that using the information content is better than just counting the path length. According to Resnik [Res99], even in the case of an information content measure, word senses have to be considered since two words from the slang vocabulary can be wrongly considered similar.

Similarity measures have also been explored between phrases or blocks of phrases. Hearst [Hea94] identifies lexical cohesion relations between pseudo-sentences of certain length by using a cosine measure and taking into account the frequency of terms in each block of sentences. Metzler et al. [MBC+05] have explored sentence-to-sentence similarity in an attempt to discover the original source of a document. They define five similarity levels; "unrelated", "on the general topic", "on the specific topic", "same facts" and "copied" and they apply similarity measures such as word overlap, frequency measures and probabilistic ones. In their initial experiments the word-overlap seems to outperform.

Evaluation. For the purpose of evaluating the similarity measures proposed, most researchers compare their similarity scores with the human judgment scores. The closer the scores are to the human results, the better the measure is. Varelas et al. [VVR+05], as well as, Seco et al [SVH04] calculate the correlation between the similarity scores obtained by their measures and the human scores gathered by the experiment of Miller and Charles [MC91].

Resnik [Res99] replicates the experiment of Miller and Charles using the same nouns they had used. Budanitsky and Hirst [BH01] agree that comparing against human answers is the best way but they point out that the human judgments consist of a small set of answers that reflect the tendency of the users to give the most dominant sense to a word.

2.6 Opinion Mining

Opinion Mining is a part of Text Mining that concentrates more on the opinions that are expressed within a text. It concerns the mining of subjective statements from texts, the identification of opinions, the estimation of opinion orientation and the extraction of arguments that relate to opinions. The identification of the opinion polarities and strength inside a text has plenty of applications and challenges.

The mining of opinions has become significant due to the vast amount of opinion information that resides mainly in web documents. Review sites, blog repositories, web forums and chat systems contain data that can be important for any decision needed to be made by companies, customers, politicians. Product reviews are interesting not only by future customers but also by companies and marketing departments [HL04]. Tracking opinions on certain subjects may give an idea of what people feel about certain political decisions or how they react against particular events.

The identification of the opinion orientation can be manual, corpus-based and dictionary-based [AS06, Liu07]. The manual identification requires a lot of human effort and it is quite costly. The corpus-based identification considers syntactic and statistical properties such as word co-occurrence, and it faces the problem of the words being domain dependent. The dictionary-based approach uses hierarchies and ontologies such as WordNet in order to identify the sentiment orientation but this approach faces the problem of the lack of context information in these hierarchies [AS06, Liu07].

2.7 Opinion Mining Techniques

Hatzivassiloglou and McKeown [HM97] were among the first to deal with opinion classification. They focus on adjectives and they study phrases where adjectives are connected with conjunction words such as "and" or "but". They construct a log-linear regression model so as to clarify whether two adjectives have the same orientation. The accuracy of this task is declared to be 82%. Their technique is described in the following steps:

- extract from a text the conjoined adjectives that are connected with the words "and"/ "but" etc.,
- run a supervised algorithm that builds a graph where the vertices are the adjectives and the links determine same or opposite orientation,
- run a clustering algorithm to separate the graph into two classes,
- assume that the cluster with the highest frequency is the one that shows positive orientation.

Let us see, now, some of the works that followed.

Use of Seed-List. An important work in the field is that of Turney and Littman [TL03] who use a pointwise mutual information (PMI) and a latent semantic analysis (LSA) measure to find out the statistical relation between a specific word and a set of positive or negative words. They construct a seed set which contains words that can be classified as either positive or negative independently of the context e.g. "excellent" is always positive. The LSA-based measure gives better results than the PMI-based one.

Their approach can be described in the following steps:

- part-of-speech tagging in order to identify adjectives and adverbs,
- extraction of 2-word phrases where one word is an adjective or adverb,
- calculation of the association between two words w1 and w2, where w1 is a word in the review and w2 is a word from the seed set. The association is calculated by the statistical measures LSA or PMI, based on the co-occurrence of two words,

- calculation of the sum of the LSA or PMI measurement between a word and the words from the positive seed set. Subtraction of this sum from the sum of the association between the same word and the words in the negative seed set,
- classification of the review as positive or negative according to the average semantic orientation of the review phrases.

Kamps et al. [KMMR04] focus on adjectives, saying that it is the adjectives that determine the semantic orientation. They follow the same logic as the work of [Tur02] with the difference that they use WordNet to define the semantic distance between the adjectives of a text and a set of already tagged words. The distance is defined as the path length between two graph vertices which contain words. They calculate the distance of a word from both "good" and its antonym "bad" and they propose three measures; the evaluative measure that shows how "good" or "bad" a word is, the potency measure that measures how "strong" or "weak" a word is and the activity measure that points out "activeness" or "passiveness".

[Wie00] deals with the distinction between objective and subjective sentences i.e. between facts and opinions. They deal with 3 subjectivity types: positive, negative and speculation. They follow the process:

- construction of a seed set by manually tagging the subjective adjectives of a corpus and determine a strength score for each of them from 1-3,
- populate the seed set as follows: for each subjective adjective of strength 3, find 20 synonyms or near-synonyms by using a distributional similarity measure [Lin98] or WordNet,
- add semantic features to adjectives. The features are the semantic orientation and the gradability (whether a word can modify a noun or it can be used in comparative sentences).

Their results show that the probability of a sentence being subjective, given that there is at least one adjective in the sentence is 55.8%. Also, the sentences that contain an adjective that exists in the expanded seed set and the list of automatically identified positive polarity adjectives are subjective by 71%. They claim that ontologies and dictionaries are not sufficient to help distinguishing between facts and opinions because they are not tagged with subjectivity.

Constructing a seed set with the right adjectives is not a straightforward task. [HDP+08] present a work in which they generate automatically a dictionary of adjectives. Initially they collect their data by getting web documents that contain negative and positive opinions. In order to determine the orientation of the opinions they use the seed set of [Tur02]. Then, they expand the initial seed set by extracting more adjectives from the collected documents. The extracted adjectives have to be related to more than one adjective found in the initial set. Finally, each document is classified as positive or negative according to the number of positive or negative adjectives it contains. The experiments show that following this approach the seed set is expanded with relevant adjectives that help in the correct classification of documents according to the opinion polarities.

A significant work in the field of Opinion Mining is that of [HL04]. They deal with product reviews written by customers on web sites. Their objective is to produce a structured summary that informs about positive or negative statements that are made for product features. Their process is the following and the model they use is presented in [Liu07]:

- find out product features that are discussed in the reviews (e.g. camera size, camera image etc.). This is done by selecting the frequent words, assuming that people often use the same words to describe features. Label sequential rules and patterns are used,
- identify opinion sentences and their orientation. An opinion sentence is defined as a sentence that contains both a feature and one or more adjectives. They use a seed list of 30 basic adjectives. For each adjective in the reviews, they check whether it exists in the seed list or it is an antonym or synonym of a word in the seed list. Every time the orientation of an adjective is found, the seed list is expanded with this adjective,
- infrequent features are identified by looking for the nearest noun phrases to an opinion word,
- summarization of results. Each sentence is assigned the orientation of the majority of its part-orientations.

[DL07] improve the previously mentioned [HL04] system by assigning an orientation score to each opinion word found in a sentence. The score takes into account the semantic orientation of the opinion word that is located near the feature-word and the distance between the feature and the opinion word. In this way a low score is given to the opinion words that are far from the feature.

[ES05a] divide the Opinion Mining tasks in 3 categories; identification of whether a phrase is a fact or an opinion, identification of the orientation of an opinion and identification of the strength of the orientation. Their method presented in [ES05a] outperforms the results of known methods such as that of [HM97] and [Tur02]. This method is based on the assumption that terms with similar orientation tend to have similar glosses. The terms of the text are presented as vectors of glosses and they are weighted by tf-idf. They start with the seed set of [KMMR04] and [TL03] that is enriched with the use of a thesaurus. Their research has resulted in SentiWordNet [ES05b], a lexical resource like WordNet with the difference that each synset is associated to a score describing how positive, negative or objective the particular synset is. The score is the proportion of a committee of classifiers used in order to label every synset as positive, negative or objective. This resource is still under evaluation.

No Seed-List used. [PLV02] deal with sentiment classification of movie reviews. Their experiments show that algorithms such as Support Vector Machines and naïve Bayes, that give good results in topic categorization, do not perform as well during sentiment classification. Additionally, they point out that the presence or absence of a word seems to be more indicative of the content of a review rather than the frequency with which a word appears in a text.

A voting system which combines various classifiers is used in [PRDP08]. The system is classifying documents according to the opinion they contain by applying linguistic techniques, vector-space reduction methods and a voting scheme which aggregates some classifiers.

An original, very recent, way of calculating the polarity and strength of opinions that differs from the techniques mentioned, is proposed in [GIS07]. They use the feedback comments posted by users in a reputation system such as "eBay" and they calculate the effect of these comments on the prices of the products sold. The orientation and the strength included in the opinion of a user's feedback are inferred by observing the changes in the respective product prices. If, for example, a certain opinion results in the reduction of a product's price, then this opinion is considered to be negative and its strength is measured on the basis of how much the price has been reduced.

In Table 2.4, the approaches mentioned are presented in ascending order of the year they have been proposed. The majority of the mentioned approaches focus on adjectives and adverbs. They use a seed list and they attempt to find out the relation between the words that appear in a text and the words of the seed list. The difference lies in the similarity measure used to calculate the association between words. Some use WordNet, others use statistical measures. Some approaches give also importance to the percentage of how positive or negative a word is.

2.8 Conclusion

The continuous expansion of textual data has led to the need for Text and Opinion Mining techniques in order to better study, exploit and analyze the textual resources. Text and Opinion Mining are two open research fields in which the issues discussed in this Chapter are still not finalized. For the purpose of approaching these issues, it is better to clarify the mining objective before the data analysis starts, since each task has different requirements.

Taking into account the language a text is written in is important since the language highlights the morphological or syntactic analysis needed. Moreover, the domain of a text collection underlines what technical terms may be present in the text or which words are redundant. Certain decisions and approaches may not be suitable for every type of text [KU96] due to the fact that term distribution varies between abstracts, articles, and collections of articles.

Natural Language Processing interacts with Text Mining. More measurable results, though, are needed so as to conclude which NLP techniques can be applied to what Text Mining applications [KP04, KP06]. In general, we should think carefully before reducing the feature list, removing stop words or applying lemmatization techniques to the texts. Noisy data may also prevent some techniques from working efficiently, so they should be corrected before the processing starts.

Table 2.4: Opinion Mining Approaches in ascending order of the year they have been proposed.

Approach	Details
Hatzivassiloglou and Mckeown [HM97]	Using categorization to detect orientation of conjoined adjectives.
Wiebe [Wie00]	A list of adjectives is expanded by WordNet. Distinction between subjective and objective phrases.
Pang et al. [PLV02]	Absence/presence of words.
Turney and Littman [TL03]	Seed List: {good, nice, excellent, positive, fortunate, correct, superior}, and {bad, nasty, poor, negative, unfortunate, wrong, inferior}. Statistical association (PMI, LSA) of a word with the words of the seed list. Extraction of adverb, adjective phrases. LSA results were better than the PMI ones.
Hu and Liu [HL04] and later Ding and Liu [DL07]	Initial seed list of 30 adjectives expanded by WordNet. Distance of opinion words from features. Objective: orientation of product reviews.
Kamps et al. [KMMR04]	Seed List: {good, bad}. Calculation of the path length in WordNet between adjectives and seed list.
Esuli and Sebastiani [ES05a, ES05b]	Seed lists used: [KMMR04] and [TL03] lists expanded by WordNet. "SentiWordNet" scores to each WordNet synset.
Ghose et al. [GIS07]	Infer opinions regarding a product by observing the changes of its price in e-Bay.
Harb et al. [HDP+08]	Seed set of [TL03] that is expanded with adjectives collected from web documents. Work oriented on the automatic construction of an adjective-dictionary.

The ambiguity is a characteristic of free text. As a result, word sense disambiguation will need to take place during the processing of certain phrases or words that are considered important for the text semantics. Identifying collocations can also help in disambiguating the meaning of some phrases.

The representation of a text is a crucial issue. Most of the researchers agree that an extension of the bag-of-words model is essential but there is still no agreement as to which kind of text properties and features should be taken into account. The attributes of the representation model depend on what kind of information we want to capture. Background knowledge, word context, and word or phrase location can be some desired properties. The text features selected can be identified with the help of tokenization and dimension reduction techniques. It is important, though, to consider where features will be looked for since certain document sections, such as the "References", should better be avoided [YHM03]. Using a combination of words and phrases is recommended. Concepts can be part of the representation as well, but more research is required on this matter.

Classifying a text collection into categories may enable the text processing. The similarity measures chosen for the categorization depend on which type of semantic or statistic distance between documents needs to be captured. The measures can apply to words, phrases, vectors or hierarchies. A combination of both syntactic and semantic measures may be considered.

New, previously unknown knowledge can also be identified by studying the semantic relations between the information stored in databases and the existing literature. This is an open issue that can be explored with the help of Text Mining and database methodologies.

Regarding the field of Opinion Mining, there are still a lot of challenges to be met. The mining of forums and online discussions is a challenge on its own because it has some particular characteristics that are issued from the nature of the forum itself. The use of colloquial language, the fact that the comments are entered by several people at different times, with different intentions and various opinions, show the difference in the analysis needs of a forum compared to a text that appears in an online newspaper.

A very interesting issue is to monitor through the texts how opinion changes over time. This will allow observing whether a product improves as the time passes, whether people become more satisfied with certain provided services, or even whether people are finally convinced and change their opinions after a long discussion in a forum.

The existence of sarcastic and ironic statements in a text cannot be identified with the current approaches. This could lead to erroneous orientation assignment and misleading opinion mining. [TL03] highlight the importance of context, since a positive word may have a negative meaning in a metaphorical or ironic context.

In this thesis we focus on the representation and mining of online discussions due to the various challenges issued by this type of text. We use the knowledge provided by the aforementioned Text and Opinion techniques and we propose a discussion model which enables the identification and mining of the parts of the discussions that may hold interesting information.

Chapter 3

A Framework for Online Discussion Analysis

Online discussions have recently been analyzed from a user-oriented point of view. They have been regarded as social networks and they have been represented by user-based graphs whose vertices represent discussion participants. In this Chapter, we present a new framework for discussion analysis. We focus on the structure and the opinion content of the discussion postings and we are looking at the social network that is developed from a semantic point of view. We formally define a model whose purpose is to provide complementary information to the knowledge extracted by the social network model. We present the measures that can be defined based on the new model, and we discuss how these measures facilitate the analysis of an online discussion. We apply these measures on real web forums and we explain how the inherent structure of the model reveals useful information about the forum itself.

3.1 Introduction

The development of Web 2.0 has resulted in the generation of a vast amount of blog repositories, review sites, web forums and online discussions. In this type of discussions people express opinions, criticize products and ideas, exchange knowledge and beliefs. Tracking opinions on specific subjects allows the identification of user expectations and needs, feelings of people about certain political decisions or reactions against particular events. As a result, mining and analyzing information that resides in online discussions is significant.

An online discussion can be represented by a graph where the vertices are knowledge entities (users, messages etc.) and the edges between them show relationships. Hence, a discussion can be analyzed by techniques of the Social Network Analysis (SNA) which is the mapping of relationships between people, organizations or other information/knowledge processing entities [HLL07].

Most existing works view an online discussion as a network in which users meet and contact each other, form communities and acquire certain roles. Online discussions are usually modeled by a social network in the form of user-based graphs whose vertices represent users that are connected with each other according to who speaks to whom. Such graphs are analyzed by social network techniques [CSW05].

The application of the social network model to an online discussion provides information about how the users interact with each other. The opinion information contained in the discussion as well as the discussion structure is lost. By taking into account the structure of the postings and their opinion content, we can become more familiar with the users and get to know better their attitude during the discussion. We can observe whether there is an important opinion presence in the discussion and if so, we can measure its amount.

In this Chapter we describe a new framework for discussion analysis by combining social network and Opinion Mining techniques. Opinion Mining is the field that deals with the opinion identification inside texts. An overview of this field has been given in the previous Chapter.

Apart from the work presented here, there are no existing techniques that explicitly combine these two fields. Our objective is to study the structure of an online debate that takes part in a well-defined domain and analyze the user reactions, preferences, and opinions on a certain subject.

The contributions presented in this Chapter are summarized in the following:

1. We propose a framework for analyzing online discussions: the structure of the discussion is seen from the point of view of exchanged messages rather than of the users who participated in the discussion. Currently most online discussions are seen from a user-oriented rather than a content-oriented point of view.

2. A combination of Social Network Analysis, Text and Opinion Mining techniques: application of topic and opinion knowledge to the postings,

identification of the relationships between them and representation of a discussion by the proposed model.

3. Definition of novel measures that are based on this model. These measures use structural, opinion, temporal and topic information in order to facilitate the analysis of a discussion.

The proposed framework allows the identification of the sentiment flow in a discussion as well as the mining of the discussion parts that contain opinions. It enables the acquisition of a content-oriented view of the discussion and the focus on the parts that have caused reactions by the participants. It gives an indication of the variety of opinions received through reply postings by a message and it monitors the opinion behavior of a user and towards a user during the discussion.

The objective of the proposed model is not to replace the social network represented by user-based graphs, but to provide additional, complementary information. It could be used together with the user-based graphs in order to enrich and better handle the knowledge extracted from a discussion.

This Chapter is structured as follows. Section 3.2 discusses existing research in discussion analysis and graph-modeled social networks. Section 3.3 describes the kind of discussions we are dealing with and the motivation of our work. Section 3.4 presents the Post-Reply Opinion Graph. Section 3.5 continues by defining the properties of the novel model together with its basic components. Section 3.6 defines measures such as structure, opinion, temporal and topic-oriented ones. In Section 3.7 and 3.8 we use the new model and the defined measures in order to analyze two discussions; an artificial and a real web discussion respectively. Section 3.9 presents the contribution of the new model by highlighting the main differences between the Post-Reply Opinion Graph and the user-based one. Finally, Section 3.10 concludes.

3.2 Social Networks and Online Discussion Analysis

The Social Network Analysis deals with the analysis of the relationships that exist between entities in a social network. For instance, in a social network of people, the analysis can include who is friend with whom, who can influence which group of people, whom can have access to the information that goes through the network etc.

Lately there has been a growing interest in this field, especially as to how it gets involved with knowledge discovery and data/web mining. For instance, analyzing the behavior of users in online discussions or discover how users form communities and are affected by them are interesting works.

Most research regarding discussion analysis focuses on analyzing the interaction between users or discovering how users form communities and are affected by them. Until now we have seen no works that examine automatically how the opinion content appears, influences and flows within a network of messages. Nevertheless, our work has been influenced by existing research and works in the social network domain.

One of these works is that of [HLL07] that analyzes the Innovation Jam 2006 among IBM employees and external contributors. The representation of the discussion is seen from the point of view of postings rather than users. The difference from our work is that our objective is not to find out the degree of innovation of a discussion but to identify the opinion flow in it. In [HLL07] they do not consider the opinion content of the discussion. Moreover, in our case, the participants of the discussion come from different backgrounds as opposed to the Innovation Jam - so they have different concepts and beliefs. Also, while in the IBM Innovation Jam the users are known since they are specific IBM employees, in the discussions we analyze users remain anonymous. Anonymous users tend to express more freely their opinions.

In [MCD08] they have analyzed discussions in the form of forums in the domain of tourism and they have extracted information regarding user sentiments and tourist destinations. They apply syntactic and semantic processing techniques and they adapt the grammar rules or the opinion words they try to identify according to the domain. They do not, though, represent the discussion as a graph.

Discussion analysis has also been dealt with in [ZAA07]. They analyze the Java Forum by using Social Network Analysis methods for the purpose of automatically identifying user expertise. They represent the social network of the forum with a graph whose vertices represent users. Their objective is different from ours since we concentrate on the content rather than the participants of a discussion and we do not seek to find experts.

A work with the objective of separating a set of newsgroup users in those that are for or against a topic is presented in [ARSX03]. In this work they represent a newsgroup as a user-based graph and they base their analysis on the “reply-to” links between the users. They focus on the users and not on the postings and although they consider the presence of agreement and disagreement, they do not actually take the opinion polarities into account.

Roles are assigned to user vertices of a graph in [FSW06] and [STE07]. We have been inspired by these works in the sense that each vertex is different and its position in the network carries information about how it affects the rest of the network. Both works, though, differ from ours both in the representation and the objective aspect. [FSW06] analyze newsgroups by applying social network techniques and they interpret online communities by assigning roles to the members of the groups. This is done by observing how people relate to each other in a graph-based model of post-reply relations. They notice that short discussion threads point out question-answer exchanges and longer threads indicate proper discussions.

[STE07] introduce a new measure that defines the number of communities to which a vertex is attached. Using this measure they assign roles to vertices by considering the community structure in the network of the vertex. Defining roles in this way, improves the performance of link-based classification and influence maximization tasks.

3.3 Problem Definition and Motivation

In this thesis, we deal with web discussions which are characterized by the following features:

- **Subject.** Each discussion has a specific subject which is usually indicated by a title that appears on top of the web page of the online discussion. An example of a subject may be: "what do you think about the economic crisis in France?".

- **Users.** In each discussion users identify themselves with a user name in the form of a "pseudo".

- **Postings.** Each user can participate in the discussion by posting messages. A user can either write a new message or reply to an already posted one.

- **Relations.** A posting A that replies to an existing posting B is related to it by a "reply-to" relation. Some but not all web sites of online discussions track such relations. In our work, we consider these relations to be known.

An extract of one such discussion is depicted in Figure 3.1. In this Figure, the discussion is taken from the site http://www.huffingtonpost.com/ and its *subject* is "Banks Prevailing In Tug Of War Over Stress Test Results". The *users* are identified on top of each message by a pseudoname (here it is replaced by a white square) and the *postings* follow. The reply-to *relations* are obvious through the indentation presentation scheme of the site.

Figure 3.1: An online discussion as it appears in the http://www.huffingtonpost.com/ web site. The pseudonames are removed in order to cater for privacy.

In current research, such online discussions are usually modeled by a social network in the form of user-based graphs. We believe that an online discussion forms a type of a social network but it is not a social network itself. One difference between online discussions and standard social network interactions is the relationships between the participants. In discussions, the participants do not necessarily know each other and they do not necessarily intend to develop a relationship between them.

A discussion participant may post a follow-up message without even noticing the person to whom the message intends [FSW06]. The reason is that in an online discussion users do not personally know the participants. They reply to a message and not to a user. They reply because they are interested in the discussed topic and not because a specific person has spoken [ZAA07]. They do not mind about who the user is but about what s/he says.

The importance of online discussions lies in their content, since the posted messages may include opinions and criticism on certain ideas and beliefs. They may contain product reviews or general knowledge about current events. The current social network models do not allow focusing on the content of a discussion.

In conclusion, another type of modeling is required that not only takes advantage of the social network models but it also exploits the content and offers additional knowledge regarding the discussion. Thus, our objective is to represent an online discussion in such a way so as to exploit the knowledge about its structure. In addition, we are interested in populating the model with opinion data extracted from the postings of the discussion.

3.4 Post-Reply Opinion Graph

In this Section we present a novel approach for representing web discussions. The new representation allows us to exploit the structural characteristics of a discussion and analyze it from a semantic and opinion oriented point of view.

The online discussions are characterized by a number of postings sent by users who have registered to the discussion site with a username. The postings are either a reply to an already existing message or they are sent as a new, independent message that signals the beginning of a new discussion thread. The relations "reply-to" between the exchanged messages of a discussion point out which message replies to what and they are considered to be known.

The model we propose for the representation of web discussions is based on graphs. Most graph-based existing approaches consider users to be the vertices of the graph. In our model, we propose to use message objects as the vertices. A message object represents a posting that has been sent to the discussion and it encapsulates information such as the content of the message and the author who has written it. Hence, we represent the discussions by a directed graph $G = (V, E)$, where V is the set of vertices which denote message objects and E are the edges that show the relations "reply-to". We call this graph a Post-Reply Opinion Graph and we define it as follows:

Definition 1. *A **Post-Reply Opinion Graph** (PROG) is a directed graph $G = (V, E)$ with a vertex set V and an edge set E. Each vertex represents a posting and each edge $e_{v'v} = (v', v)$ points out a reply direction from the vertex v' to the vertex v. A posting v represented by a vertex is defined as:*

$$v = (m_v,\ op_v,\ u_v,\ tm_v),$$

where m_v is the actual content of the message, op_v the opinion polarity included in the message, u_v the user that has written it, and tm_v the timestamp that shows when the message was posted.

In Figure 3.2 we can see an example of a Post-Reply Opinion Graph.

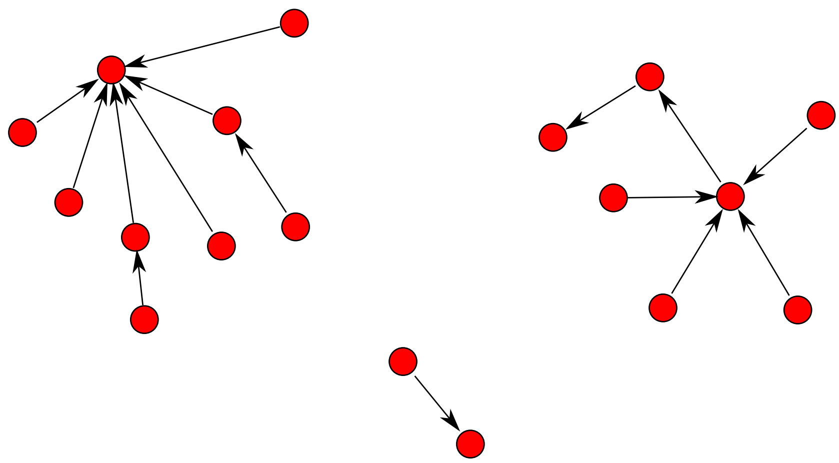


Figure 3.2: A Post-Reply Opinion Graph of an online discussion.

One main characteristic present in the definition of a Post-Reply Opinion Graph is the *opinion* op_v of a vertex v of the graph.

Definition 2. *We define the opinion of a posting represented by the vertex v as:*

$$op_v = polarity_v,$$

where $polarity_v$ *takes values in* $\{n, o, p\}$ *if the opinion expressed in the posting* v *is negative, objective (i.e. no opinion) or positive respectively.*

The *opinion* op_v captures the opinion polarity expressed in the message m_v. It may be negative (n) or positive (p). In the case where the content of the message is objective, we consider that there is no opinion included (o). The *polarity* is calculated by Opinion Mining techniques such as those mentioned in Chapter 2, for instance techniques presented in [DL07, GIS07, HM97, HL04, TL03]. Applying Opinion Mining methodologies is out of the scope of this thesis and this is why we will consider that the *opinion* expressed by a vertex is known.

The *author* of the message u_v is encapsulated in the message object. In this way, information about the author is not lost. As a result, the social network of users can be extracted from the proposed model. This is an important property of the Post-Reply Opinion Graph, since the information provided by the social networks can still be exploited.

The notion of *time* is also encapsulated in the proposed model. As a result, the future and the past of a vertex can be easily traced. The direct past of a vertex v, is one and only one message object that has taken place immediately before the message object represented by this vertex v. Similarly, the future of the vertex v, $\{v' \in V : (v', v) \in E\}$ contains message objects that have been posted after the posting represented by v.

Apart from the implicit temporal information, the proposed graph includes the information of time explicitly in each vertex. The chronology of each posting is captured through the tm_v timestamp. Knowing the time of having posted a message allows us to enrich the structure of the graph. We can understand this concept through a small example.

Let us assume that in a certain discussion, the posting v_3 sent at time tm_3 replies to the posting v_1 posted at time tm_1. In the same discussion, the posting v_2 posted at tm_2 replies to no message and it was sent after the posting v_1 and before the v_3 i.e. $tm_1 < tm_2 < tm_3$.

There are three cases regarding what the message v_2 may be. It could be:

1. a message that has actually been influenced by existing messages but does not explicitly reply to one of them,

2. a posting that refers to a new argument or a new topic that has not been mentioned in previous postings and

3. noise in the form of a spam or irrelevant to the subject message.

In the first case, and as it can be seen from Figure 3.3, the graph can be enriched with the knowledge of the chronology. On the right hand side of the Figure, the chronology edges are shown with dotted lines. This can help us to better identify which postings may have influenced a message not just by exploring the structure of the discussion, but also by analyzing the chronology links. In the example shown in Figure 3.3, the author of v_3 has replied to v_1, but it may be after s/he has read the posting v_2. Exploiting the chronology links is a future issue in the Post-Reply Opinion Graphs.

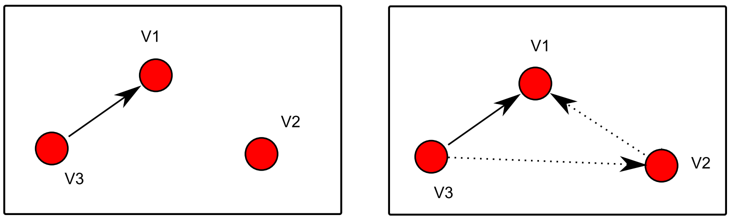


Figure 3.3: The chronology knowledge enriches the structure of a discussion. In the Figure, v_2 has been sent after v_1 and before v_3, while v_3 replies directly to v_1.

The second as well as the third cases can only be identified through content analysis, topic-identification and Text Mining techniques mentioned in Chapter 2.

3.5 Model Properties

Let us consider a PROG graph $G = (V, E)$ and one of its vertices $v \in V$. The number of edges adjacent from that vertex v is called the *outDegree* of v. The *outVertices* of a vertex v and the *outDegree* are defined as:

$$outVertices(v) = \{v' \in V : (v, v') \in E\}, \tag{3.1}$$

and

$$outDegree(v) = \mid outVertices(v) \mid . \tag{3.2}$$

Lemma 1. *The outDegree of the vertices of a PROG graph G is 0 or 1.*

This lemma stands because by definition, in the online discussions we model, the user can post a message that replies to maximum one existing posting.

Another characteristic of PROG graphs is that a *walk* between two vertices does not repeat any edge or vertex. Thus, we have the following theorem:

Theorem 1. *In a PROG graph G, all possible walks between two vertices are paths.*

Proof. Suppose the PROG graph G has a walk which is not a path. This means that there is a walk that repeats at least one vertex of the graph. Let this repeated vertex be v. In order for this vertex to be repeated in the walk, one of its "past" vertices v' should be adjacent to it so that an edge $e_{v'v} = (v', v)$ exists.

Since the vertex v' has happened before v, we have $tm_v > tm_{v'}$. Thus, there cannot be an edge $e_{v'v}$ because the posting represented by v' cannot reply to the posting represented by v since v happened after v'. Hence, no such walk exists that is not a path. □

Product of the previous theorem is the insight that a PROG graph cannot contain cycles. Thus, it is a forest since it contains no cycles and it can be consisted of many components. Therefore, we have the following Lemma:

Lemma 2. *A Post-Reply Opinion Graph is a forest.*

The novel graph is consisted of components whose identification allows us to define measures in order to extract useful information from such graphs. There are two basic components; the **discussion threads** and the **discussion chains**. The distinction between a discussion thread and a discussion chain becomes apparent from Figure 3.4 that shows a graph consisted of two discussion threads.

Definition 3. *The set of the* ***discussion threads*** *in a Post-Reply Opinion Graph G is the union of all the maximal connected components of G.*

The discussion threads can be "queried" either by a message m or a user u. For example, the threads where the user u has participated can be found by tracing the vertices of each thread of the graph until a message object $v = (m_v, op_v, u, tm_v)$ is found.

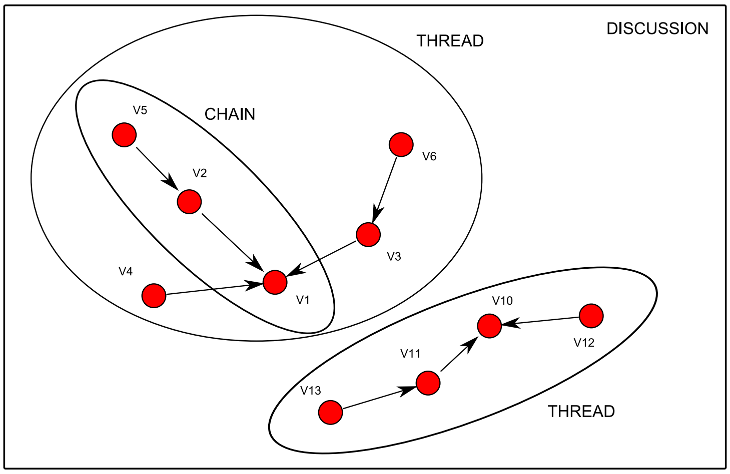


Figure 3.4: Discussion threads and chains of a discussion.

The discussion chains consist of the paths in the graph whose starting vertex is a root and ending vertex is a leaf when we inverse the direction of the edges. In order to define a discussion chain, we consider $root(G)$ to be the set of vertices of the graph G which represent message objects that do not reply to another message.

Moreover, $inVertices(v)$ describes the vertices (head endpoints) which are adjacent to the vertex v. According to the theory of graphs, we have:

$$inVertices(v) = \{v' \in V : (v', v) \in E\}. \tag{3.3}$$

The *inDegree* of a vertex shows how many reactions have been caused by the posting represented by the vertex $v \in V$. We define the *inDegree* as:

$$inDegree(v) = \mid inVertices(v) \mid . \tag{3.4}$$

A formal definition of a discussion chain follows:

Definition 4. *We define a* ***discussion chain*** *in the graph* $G = (V, E)$ *as the subgraph*

$$G_c = (V_c, E_c)$$

where
$V_c = \{v_i, v_{i-1}, v_{i-2}..., v_{i-x}\}$, $v_i \in root(G)$, $inVertices(v_{i-x}) = \emptyset$, $v_i \neq v_{i-x}$,

$v_{i-k} \in inVertices(v_{i-(k-1)}), \forall k, k \in [1, x]$ *and*
$E_c = (V_c)^2 \cap E$.

Similarly to the discussion threads, the discussion chains can also be queried by a specific message or user. The discussion chains where a message m appears are all the chains G_c of the graph G for which $\{\exists v \in V_c : v = (m, op_v, u_v, tm_v)\}$. Similarly, the chains where the user u has participated are the chains G_c of the graph G for which $\{\exists v \in V_c : v = (m_v, op_v, u, tm_v)\}$.

In Figure 3.4, the first thread is consisted of 3 discussion chains: $\{v1, v3, v6\}$, $\{v1, v4\}$, $\{v1, v2, v5\}$. The second thread is consisted of 2 discussion chains: $\{v10, v11, v13\}$, $\{v10, v12\}$.

The chains are important in a PROG graph. The longest discussion chain can point out the longest exchange of messages in a discussion. It can be seen from the point of view of the number of postings involved, so it can be measured by the maximum number of edges that start from a leaf vertex and end up to a root vertex. It can also be seen from the temporal point of view and it can be measured by the maximum time distance between a leaf and a root vertex, when we inverse the direction of the edges.

If we have more than one chain in the graph, then there is at least one vertex v that has received more than one reply. Additionally, if there exists a vertex $v \in V$ for which its reply has received another reply, then we assume that we have a generation of possible sub-discussions that start from v. Otherwise we consider to have only reactions and not sub-discussions starting from v.

For example, in Figure 3.5 we can assume that the root of the graph (which, in this case, it is only one when we inverse the direction of the edges) has caused the generation of two different sub-discussions, one of which is initiated by the vertex in black. This black vertex, in turn, is dividing the discussion into two parts. The light grey vertex has caused four reactions that have not moved the discussion forward, so we cannot assume that there are four different arguments or sub-discussions that have been invoked.

Moreover, the number of discussion chains together with the number of messages per chain, characterize the discussion. For example, in Figure 3.6, both message-graphs contain 4 discussion chains but the structure is completely different. The graph on the left-hand side contains 2 messages per chain and the focus is only on one message (the root one). The other graph contains 3 messages per chain and the focus is well distributed.

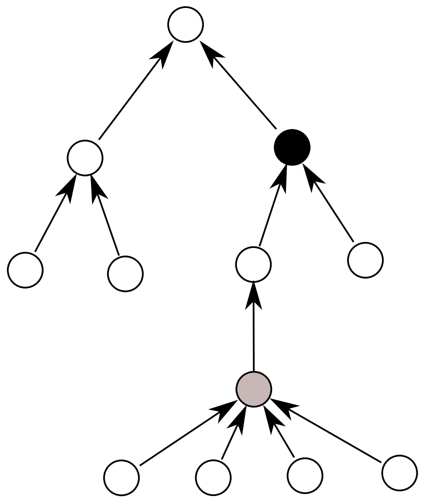


Figure 3.5: A discussion thread in which the black vertex may have possibly caused sub-discussions while the light grey one has just caused reactions.

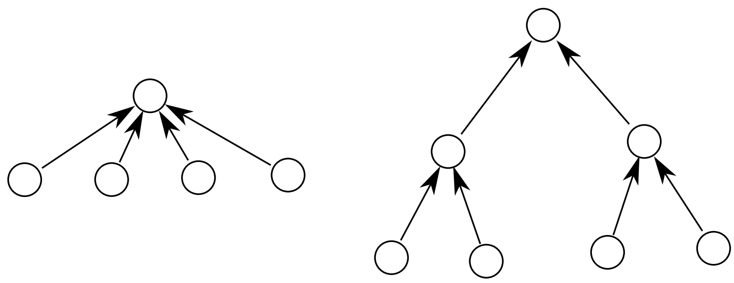


Figure 3.6: Two discussion threads which contain the same number of discussion chains do not necessarily have the same structure.

3.6 Measures based on the model

A Post-Reply Opinion Graph models an online discussion and a vertex of a PROG graph represents a posting of the particular discussion. Each vertex encapsulates information that permits the deeper investigation and analysis of each discussion.

In this Section we will consider a PROG graph $G = (V, E)$. We will look at the information held by each vertex $v \in V$ separately and we will define measures that allow us to navigate inside an online discussion and analyze it efficiently. A summary of all the proposed measures structured in tables is presented in the Appendix A.

3.6.1 Structure-oriented Measures

The PROG graph, by its definition, is not *complete* and it does not have the presence of *cycles* or *cliques* in it. Thus, theorems and measures that are defined in the theory of graphs regarding these concepts, cannot be applied in our case. Nevertheless, there are some elementary concepts of graph theory that are used in PROG graphs. These concepts are presented in this Section.

`a) Root and Leaf.` A PROG graph is a forest whose components have roots and leaves.

Definition 5. *A **root** vertex of the PROG graph is a vertex that begins a part of the discussion, while a **leaf** vertex is a vertex that ends it.*

A root vertex which initiates a part of the discussion can be interpreted in two ways:

1. Root is considered to be the vertex that initiates a discussion thread. As previously mentioned, the *outDegree* for a vertex of a PROG graph is either 0 or 1.

   Messages that signal the beginning of a new discussion thread, even if this thread is consisted of only one vertex, are considered to be "root" messages. The root messages of a graph $G = (V, E)$ are defined as:

$$root(G) = \{v \in V : outDegree(v) = 0\} \tag{3.5}$$

2. Root is considered to be the vertex that initiates a topic in the discussion. The topics can be extracted through topic-identification techniques such as those mentioned in Chapter 2. In this case, the definition becomes:

$$root(G) = \{v \in V : v \in initialVertex(G, T_i)\}, \tag{3.6}$$

for all topics T_i that have been found in the discussion.

The $initialVertex(G, T)$ is the set of vertices that initiate a topic T. It is a set and not a single vertex because there is a possibility that two or more new postings that belong to the same topic are sent at the same time:

$$initialVertex(G, T) = \{v \in V : tm_v = min_{i \in msgs(T)} tm_i\}, \tag{3.7}$$

where $i = (m_i, op_i, u_i, tm_i)$ and $msgs(T)$ are all those vertices that belong to the topic T.

The messages that end a discussion chain are called "leaf" messages and they are characterized by the fact that they have received no reply. Similarly to the "root" messages we can have two types of leaf messages:

1. Leaf is considered to be a vertex that ends a discussion thread. Hence, leaf is a message that belongs to the set:

$$leaf(G) = \{v \in V : inDegree(v) = 0\} \tag{3.8}$$

2. Leaf is considered to be a vertex that is the last one that belongs to a particular topic T in the discussion. Again, here, the topics are extracted from a discussion through topic-identification techniques. In this case, the definition becomes:

$$leaf(G) = \{v \in V : v \in finalVertex(G, T_i)\}, \tag{3.9}$$

for all topics T_i that belong to the discussion.

The $finalVertex(G, T)$ is the set of vertices that end a topic T. It is a set and not a single vertex because there is a possibility that two or more new postings that finalize the same topic are sent at the same time:

$$finalVertex(G, T) = \{v \in V : tm_v = max_{i \in msgs(T)} tm_i\}, \qquad (3.10)$$

where $i = (m_i, op_i, u_i, tm_i)$ and $msgs(T)$ are all those vertices that belong to the topic T.

By the "leaf" and the "root" definitions, we can realize that a posting which replies to no existing posting and has received no reply postings is seen as both a "leaf" and a "root".

`b) Popularity.` A concept of the graph theory is the popularity of a vertex which is defined as follows:

Definition 6. *A vertex $v \in V$ is considered to be **popular** if it has caused many reactions. The more the reactions, the higher is the popularity.*

This definition can be interpreted in various ways and it can be measured accordingly.

1. One way to measure popularity is to consider as "reactions" the direct future of the vertex $v \in V$. These are the reply vertices $inVertices$ towards this vertex. The number of edges that connect the reply vertices to the vertex v is the $inDegree$ of a vertex and it consists of one way to measure the popularity of a vertex. An example of how we measure this type of popularity is given in Figure 3.7, where the black vertex v has a popularity of 2.

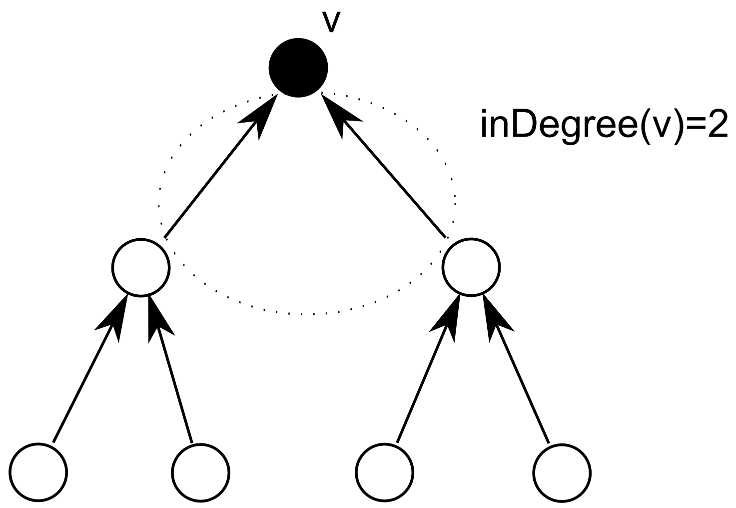


Figure 3.7: The popularity of the vertex v is $inDegree(v) = 2$.

2. Another way to measure the popularity of a vertex is to consider as "reactions" not only the direct future but also the indirect. Hence, a vertex will be considered popular if it has caused reactions and these reactions have caused reactions as well.

   Thus, we have the measure of $inVerticesExtra$ for a vertex v as:

$$inVerticesExtra(v) = inVertices(v) \; \cup \; (\cup \, inVerticesExtra(i)), \tag{3.11}$$

   where $i \in inVertices(v)$.

   The $inDegreeExtra$ is the number of the set of direct and indirect future vertices together and it is given as:

$$inDegreeExtra(v) = \mid inVerticesExtra(v) \mid . \tag{3.12}$$

   The $inDegreeExtra$ is another way to measure the popularity of a vertex. An example of how we measure this type of popularity is given in Figure 3.7, where the black vertex v has a popularity of 6.

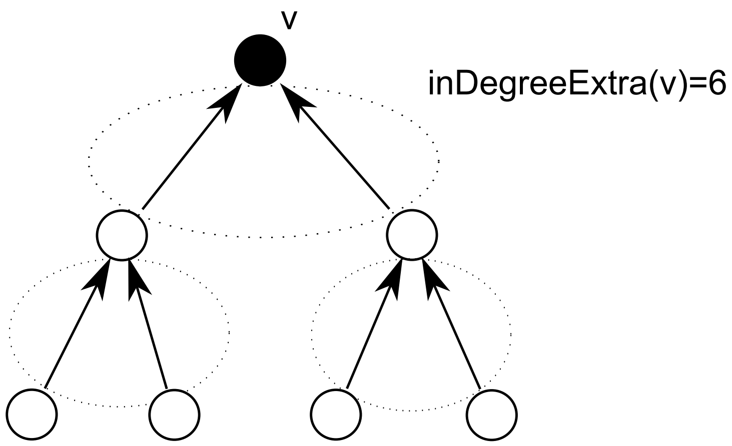


Figure 3.8: The popularity of the vertex v is $inDegreeExtra(v) = 6$.

`c) Order.` The "order" of a graph G permits to have an exact idea about its size. The components of the graph i.e. the discussion threads and chains have their own order as well.

1. Order of the graph G.

   According to the graph theory, the order of a graph equals the number of existing vertices, so we have:

   $$order(G) = \mid V \mid \tag{3.13}$$

   In order to measure the number of the discussion threads which are present in a Post-Reply Opinion Graph G, we define:

   $$orderThread(G) = \mid \{G_{thr} \in G\} \mid, \tag{3.14}$$

   where $G_{thr} = (V_{thr}, E_{thr})$ represents any discussion thread of the graph G.

   Another measure is the number of the discussion chains which are present in a Post-Reply Opinion Graph G. This measure is defined as:

   $$orderChain(G) = \mid \{G_c \in G\} \mid, \tag{3.15}$$

   where $G_c = (V_c, E_c)$ represents any discussion chain of the graph G.

2. Order of a discussion thread.

   Each discussion thread consists of a number of vertices. The order of the discussion thread G_{thr} is defined as:

   $$orderOfThread(G_{thr}) =\mid V_{thr} \mid,$$

   where $G_{thr} = (V_{thr}, E_{thr})$ is the subgraph of G that represents the discussion thread.

   Similarly we can have a measure of the number of discussion chains that belong to a particular thread:

   $$orderThrChain(G_{thr}) =\mid \{G_c \in G\} \mid, \tag{3.16}$$

   where $G_c = (V_c, E_c)$ represents any discussion chain of the subgraph $G_{thr} = (V_{thr}, E_{thr})$.

3. Order of a discussion chain.

   The order of a discussion chain is simply the number of vertices it contains and it is defined as:

   $$orderOfChain(G_c) =\mid V_c \mid,$$

   where $G_c = (V_c, E_c)$ represents a discussion chain of the graph G.

3.6.2 Opinion-oriented Measures

In this Section, we concentrate on the measures that enable us to determine the flow of the opinion inside a discussion as well as the opinion status of the participants. We introduce some concepts that allow us to define opinion-oriented measures in order to satisfy the ideas behind each concept.

`a) Opinion Status of a User.` A user may own more than one posting inside a discussion by replying, for example, to messages s/he has already received. The **message objects** that are generated by a certain user u in the whole discussion represented by the graph $G = (V, E)$ are given by:

$$msgs(u) = \{v \in V : v = (m_v, op_v, u, tm_v)\} \tag{3.17}$$

Definition 7. *The **opinion status** of a user is defined by the average opinion polarity s/he expresses inside the discussion.*

The opinion status of a user can be measured per discussion chain, per discussion thread and per discussion as a whole.

1. Opinion status of a user u in a discussion chain.

   By capturing the average opinion expressed by a user u inside a discussion chain G_c, we identify the average opinion reaction of the specific user within a specific sub-discussion. We define this concept by the following measure:

   if $\mid msgs(u) \cap V_c \mid > 0$, then

   $$avgOpFromUsr(G_c, u) = \frac{\sum_i op_{v_i}}{\mid msgs(u) \cap V_c \mid} \tag{3.18}$$

   where $v_i \in msgs(u) \cap V_c$.

2. Opinion status of a user u in a discussion thread.

   Similarly to the opinion status of a user in a discussion chain, we measure the opinion inside a discussion thread G_{thr} as:

   if $\mid msgs(u) \cap V_{thr} \mid > 0$, then

   $$avgOpFromUsr(G_{thr}, u) = \frac{\sum_i op_{v_i}}{\mid msgs(u) \cap V_{thr} \mid} \tag{3.19}$$

   where $v_i \in msgs(u) \cap V_{thr}$.

3. Opinion status of a user u in the discussion.

   The opinion of a user can also be seen globally for the whole of the discussion. In this way, we can observe users that keep a negative or positive position throughout the discussion or we can identify tendencies such as whether people tend to write more when they are unhappy or when they are satisfied with a certain situation. We define the average opinion expressed by a user u during the discussion as:

   $$avgOpFromUsr(u) = \frac{\sum_i op_{v_i}}{\mid msgs(u) \mid} \tag{3.20}$$

   where $v_i \in msgs(u)$ and $msgs(u)$ is given by the equation 3.17, `b) Opinion Reactions Towards a User.` A posting written by a user may be replied to many times or it may be ignored. For the purpose of identifying the opinion status of other users towards a particular user u we define the following measures:

1. Opinion reactions towards a user u inside a discussion chain.

   The average opinion expressed towards a user inside a chain as:

$$avgOpToUsr(G_c, u) = \frac{\sum_i op_{v'_i}}{\sum \mid inVerticesV(v) \mid}, \tag{3.21}$$

   where $v \in msgs(u) \cap V_c, inVerticesV(v) = inVertices(v) \cap V_c$ and $v'_i \in inVerticesV(v)$.

   This measure describes on average the opinion expressed in the reactions towards the postings of the specific user, within a sub-discussion. The pre-requisite is that at least one of the postings of the user has been replied to.

2. Opinion reactions towards a user u inside a discussion thread.

   Similarly to the measure for the discussion chain, we have the average opinion expressed towards a user inside a discussion thread as:

$$avgOpToUsr(G_{thr}, u) = \frac{\sum_i op_{v'_i}}{\sum \mid inVertices(v) \mid}, \tag{3.22}$$

   where $v \in msgs(u) \cap V_{thr}$ and $v'_i \in inVertices(v)$.

   This measure describes on average the opinion expressed in the reactions towards the postings of the specific user, within a thread. Again the pre-requisite is that at least one of the postings of the user has been replied to.

3. Opinion reactions towards a user u inside the discussion.

   The average opinion expressed towards a user u (having received at least one answer) during the discussion is given by the following measure:

$$avgOpToUsr(u) = \frac{\sum_i op_{v'_i}}{\sum_j inDegree(v_j)} \tag{3.23}$$

   where $v_j \in msgs(u), v'_i \in inVertices(v)$.

`c) Opinion Reactions Towards a Posting.` The opinion polarity expressed towards a specific posting can be measured in many ways:

1. Number of replies towards the posting according to their opinion polarity.

   A message object $v \in V$ may be replied to during a discussion through postings. These postings may contain objective information or they may include the sentiments of the author expressed by positive or negative opinions.

   We can always distinguish between the different postings according to their opinion polarity. The number of positive postings towards a message object $v \in V$ is defined by the number of reply vertices that contain a positive opinion. We describe the number of negative (n), objective (o) and positive (p) replies respectively as:

   $$reply(v, r) = \mid \{v' \in inVertices(v), op_{v'} = r\} \mid \tag{3.24}$$

2. The average opinion expressed towards the posting.

   Measuring the average opinion received by a message object v can give us an indication of the reactions of the participants towards the specific posting. If, for example, the average opinion is 0, this means that either the reply postings contained objective information, or there is a balance between positive and negative opinions.

   We define the average opinion received by a message object $v \in V$ that has caused reactions as:

   $$avgMsgOpinion(v) = \frac{\sum_i op_{v'_i}}{inDegree(v)}, \tag{3.25}$$

   where $v'_i \in inVertices(v)$.
   The $inDegree(v)$ points out how many replies the posting represented by the vertex v has received.

3. The variety in opinion polarity of the replies towards to the posting.

   Having described the various vertices according to the opinion polarities included in their reply postings, allows us to define a measure regarding the *opinion information* held by a vertex. We use the entropy H for this purpose, and we define the amount of opinion information held by a vertex $v \in V$ (that has been replied to), as:

$$H(v) = -\sum_{r=n,o,p} \left(\frac{reply(v,r)}{inDegree(v)} \log \frac{reply(v,r)}{inDegree(v)}\right) \tag{3.26}$$

The opinion information measured by the entropy is used in general for measuring the diversity of opinions [Kri75]. In our case it is an indication of the variety of opinions received by a vertex. If, for instance, a vertex has received reply postings that are all of the same opinion orientation, then the entropy will be 0. This may mean either that there is objective information or that there is unanimous opinion regarding the message expressed by the particular vertex.

The entropy is a measure that exploits the global (through the edges) together with the local information of the PROG graph. It is based on how the postings are linked to each other and on what opinion information they hold. A discussion analyst can distinguish the messages with high entropy among all others since these would be messages that have caused an intense discussion with various opinions.

Similarly to the *opinion information* measure per vertex $H(v)$ we defined previously, we can define the same measure per discussion chain. This measure facilitates the identification of the discussion chains that contain the maximum amount of opinion information.

Before defining the measure, let us consider a discussion chain $G_c = (V_c, E_c)$ of the graph G. We can calculate the number of vertices that point out negative (n) opinions, objective (o) statements and positive (p) opinions respectively as:

$$verticesCh(G_c, r) = \mid \{v \in V_c, op_v = r\} \mid \tag{3.27}$$

Now, the opinion information inside a discussion chain is defined as:

$$H(G_c) = -\sum_{r=n,o,p} \left(\frac{verticesCh(G_c,r)}{\mid E_c \mid} \log \frac{verticesCh(G_c,r)}{\mid E_c \mid}\right) \tag{3.28}$$

where n, o and p point out the negative, objective and positive opinion orientation respectively.

The opinion information is an indication of the variety of opinions inside a discussion chain. Similarly we can define the opinion information inside a discussion thread.

3.6.3 Time-oriented Measures

The temporal dimension can facilitate the analysis of a discussion over a time period. The temporal information is mostly exploited in relation to other information such as the topic and the opinion, and hence it is used within other measures in other sections. In this Section we present only measures which are independent from the topic and opinion knowledge.

`a) Duration of a discussion.` The duration of the discussion represented by the graph $G = (V, E)$ is apparent by looking at the temporal distance between the first v and the last v' message of the discussion. In this case, the vertex that initially appeared as a posting in a discussion is:

$$initialVertex(G) = v \in root(G), \tag{3.29}$$

where $tm_v = min_{i \in root(G)} tm_i$ and $i = (m_i, op_i, u_i, tm_i)$
and the last posting is represented by:

$$finalVertex(G) = v \in leaf(G), \tag{3.30}$$

where $tm_v = max_{i \in leaf(G)} tm_i$ and $i = (m_i, op_i, u_i, tm_i)$.
The duration of the discussion would then be:

$$duration(G) = tm_{v'} - tm_v \tag{3.31}$$

where $v = initialVertex(G)$ and $v' = finalVertex(G)$.

`b) Ancestors of a posting beyond structure.` The temporal information permits also to distinguish the messages that have been posted immediately before a specific posting. This cannot only be derived by the

$outVertices$ of a vertex. The reason is that some users choose to send messages as a new posting without replying to an existing one, even though they may have been influenced by the existing postings. The case has been explained in the "Model Properties" Section and has been depicted in Figure 3.3.

In order to see which message comes before another one we can define a measure that compares two vertices by the time of post:

$$compareTime(v, v') = \frac{tm_v - tm_{v'}}{\mid tm_v - tm_{v'} \mid}, \tag{3.32}$$

which will be -1 if $tm_v < tm_{v'}$, 1 if $tm_v > tm_{v'}$ and 0 otherwise.

Using the $compareTime$ measure we can define the ancestors of a vertex v as:

$$ancestors(v) = \{v' \in V : compareTime(v, v') = 1\} \tag{3.33}$$

3.6.4 Topic-oriented Measures

Each textual discussion message can be analyzed by Text Mining techniques such as those mentioned in Chapter 2. Therefore, a message can be tokenized accordingly and it can be cleared from the noisy data. Then, feature selection can take place so that the main keywords and the collocations are extracted. The senses of words can be disambiguated and the text can be represented by an appropriate structure. Although the application of such techniques may clarify and simplify the content of each posting, we choose not to apply them in the scope of this thesis. In any case, when we deal with large online discussions, it is more efficient to use the PROG structure in order to distinguish a subset of important messages and then apply to this subset some content-oriented techniques.

We choose to exploit the content of the different postings by using topic-identification algorithms such as [VG07]. We consider the topic per posting to be known as well as to which topics each posting belongs to.

Let the topic of a posting represented by a vertex v be denoted as $topic(v)$. Let us also consider a topic T.

Then, the postings that belong to this topic are given by:

$$msgs(T) = \{v \in V : topic(v) = T\}. \tag{3.34}$$

`a) Participation of a User in a Topic.` For the purpose of identifying whether a user has had messages inside a topic as well as the proportion of them, we define the following measure:

$$userParticip(u, T) = \frac{\mid msgs(u) \cap msgs(T) \mid}{\mid msgs(u) \mid} \tag{3.35}$$

`b) Opinion Evolution of a User in a Topic.` A significant measure is that of the opinion evolution of a person in a particular topic. This measure becomes even more important if we know or have identified the profile of the user. If, for example, the user is an expert in the domain of the discussion, then his opinion has a higher value.

This concept can be measured in various ways:

1. Average Opinion of a User in a Topic. We can approximate this measure by observing the average opinion that a user has expressed in postings that belong to a particular topic T, if $userParticip(u, T) > 0$:

$$opEvolution(u, T) = \frac{\sum_i op_{v_i}}{\mid msgs(u) \cap msgs(T) \mid} \tag{3.36}$$

   where $v_i \in msgs(u) \cap msgs(T)$.

2. Difference in opinion polarity of a User u between two timestamps of the same Topic T.

$$opEvolution(u, T, tm_v, tm_{v'}) = \mid op_v - op_{v'} \mid \tag{3.37}$$

   where $v, v' \in msgs(u) \cap msgs(T)$ and $v = (m_v, op_v, u_v, tm_v)$, $v' = (m_{v'}, op_{v'}, u_{v'}, tm_{v'})$.

   The measure will give a 0 result if the user has not changed opinion between the two instances. It will be 1 if the user has moved from or to an objective posting and it will give the value of 2 if the user has gone from a positive to a negative opinion and vice versa.

`c) Opinion expressed for a Topic.` The opinion can also be calculated on average per topic without specifying a user. Then, the average opinion polarity in the discussion for a topic T is:

$$avgOpTopic(T) = \frac{\sum_i op_{v_i}}{\mid msgs(T) \mid} \tag{3.38}$$

where $v_i \in msgs(T)$.

In the same way, we can measure the average opinion polarity for a topic inside a discussion chain G_c or a discussion thread G_{thr}. The measures are respectively:

if $\mid msgs(T) \cap V_c \mid > 0$, then

$$avgOpTopic(G_c, T) = \frac{\sum_i op_{v_i}}{\mid msgs(T) \cap V_c \mid} \tag{3.39}$$

where $v_i \in msgs(T) \cap V_c$,

and

if $\mid msgs(T) \cap V_{thr} \mid > 0$, then

$$avgOpTopic(G_{thr}, T) = \frac{\sum_i op_{v_i}}{\mid msgs(T) \cap V_{thr} \mid} \tag{3.40}$$

where $v_i \in msgs(T) \cap V_{thr}$.

`d) Topic Popularity.` The popularity of a topic can be measured inside a discussion, a discussion thread or even a discussion chain.

1. Topic Popularity in a discussion chain. A measure that calculates the proportion of the vertices of a certain topic T per discussion chain is given by the formula:

$$topicPop(G_c, T) = \frac{msgs(T) \cap V_c}{\mid V_c \mid} \tag{3.41}$$

where $G_c = (V_c, E_c)$ is a discussion chain of the graph G.

2. Topic Popularity in a discussion thread.

   Similarly, the proportion of the postings of a certain topic T per discussion thread is given by the formula:

$$topicPop(G_{thr}, T) = \frac{msgs(T) \cap V_{thr}}{\mid V_{thr} \mid} \tag{3.42}$$

   where $G_{thr} = (V_{thr}, E_{thr})$ is a discussion thread of the graph G.

3. Topic Popularity in the discussion. Similarly, we have the same measure for the discussion represented by the PROG graph $G = (V, E)$:

$$topicPop(G, T) = \frac{msgs(T)}{\mid V \mid}. \tag{3.43}$$

`e) Ancestors of a posting by topic.` Having the information of which topic a posting belongs to enables us to see the actual ancestors of a posting, not only by time (as in the time-oriented measures) but also by topic.

$$ancestors(v, T) = \{v' \in V : tm_{v'} < tm_v, topic(v') = topic(v)\} \tag{3.44}$$

where $v = (m_v, op_v, u_v, tm_v)$, $v' = (m_{v'}, op_{v'}, u_{v'}, tm_{v'})$.

This measure can facilitate the identification of links between messages that are not captured by the structure of the discussion.

`f) Descendants of a posting by topic.` Similarly to the ancestors, we can define the descendants of a posting by topic. This could be also seen as a measure of popularity for a vertex v, considering as “reactions”, the vertices that have followed in time and they also belong to the same topic T as the specific vertex v. We define the descendants of a vertex v as:

$$descendants(v) = \{v' \in V : tm_{v'} > tm_v\} \tag{3.45}$$

where $v = (m_v, op_v, u_v, tm_v)$, $v' = (m_{v'}, op_{v'}, u_{v'}, tm_{v'})$.

Using the measure of the *descendants* we define the popularity of a vertex v as:

$$inDegreeDesc(v) = \mid inVerticesExtra(v) \cup inVerticesExtra(v') \mid, \quad (3.46)$$

where $v' \in (descendants(v) \cap root(G)) \cap msgs(T)$,

and $msgs(T)$ are all those vertices that belong to the topic T.

An example of how we measure this type of popularity is given in Figure 3.9, where the black vertex v has a popularity of 4. The light grey vertices show the vertices that belong to the same topic as the black vertex.

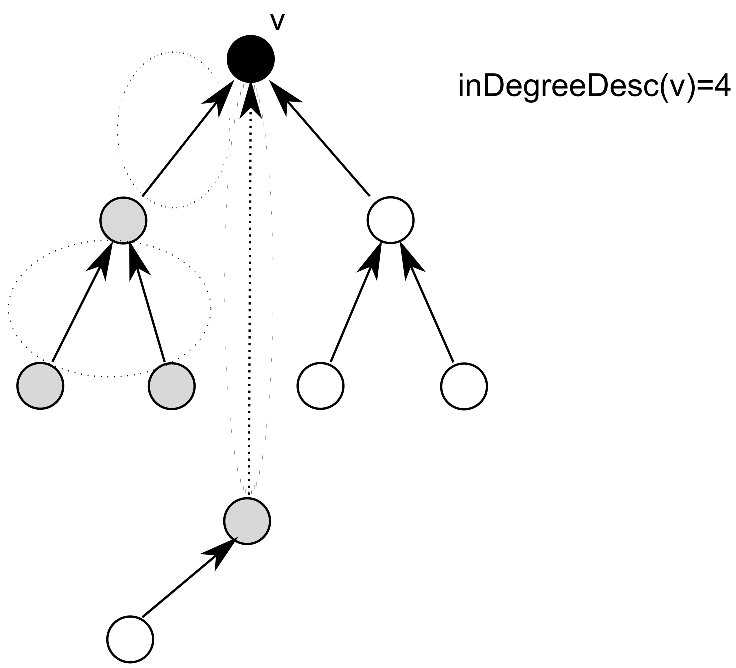


Figure 3.9: The popularity of the vertex v is $inDegreeDesc(v) = 4$. The light grey vertices belong to the same topic as the black vertex, while the white ones belong to a different topic. The dotted edge shows that the grey vertex has followed in time the black one.

3.6.5 User-oriented Measures

User-oriented measures can be more appropriately defined in a user-oriented graph that describes a social network. Such measures consist of the degree of the membership of a user in a community, the role of the user, how central s/he is in the network etc.

In the case of discussions, as we have explained in the motivation of this thesis, the users do not really form communities and networks. Thus, the existing user-oriented measures cannot be applied in PROG graphs. Nevertheless, since the information about the user who has posted a message is encapsulated in the vertex of the Post-Reply Opinion Graph, we can identify the users who are "central as people who start conversations, or as the ones who end them" [FSW06]. In [FSW06], these users are identified from a user-based graph through the *inDegree* and the *outDegree* distributions.

Therefore, the users that start a sub-discussion represented by the graph G are:

$$usersStartDisc(G) = \{u_v : v \in root(G)\} \tag{3.47}$$

while the users that end a sub-discussion are the ones in the set:

$$usersEndDisc(G) = \{u_v : v \in leaf(G)\} \tag{3.48}$$

The users that are defined by $usersStartDisc(G) \cap usersEndDisc(G)$ are the users who have written messages which started a discussion thread but no one responded to them.

More interesting measures regard the opinion status of a user inside a discussion and they have been previously described in the Opinion-oriented measures Section.

3.7 Analysis of an artificial discussion

Let us present a small artificial discussion in Figure 3.10. The indentation shows "reply-to" links. Our objective is to exploit the inherent structure of such a discussion and facilitate its mining by using some of the proposed measures.

In this discussion, we have six participants (A, B, C, D, E, F) who send postings regarding holiday destinations. The exchanged messages are 15. The message flow in ascending temporal order is shown in Table 3.1. Information about the topic is not included in this example in order to keep it simple.

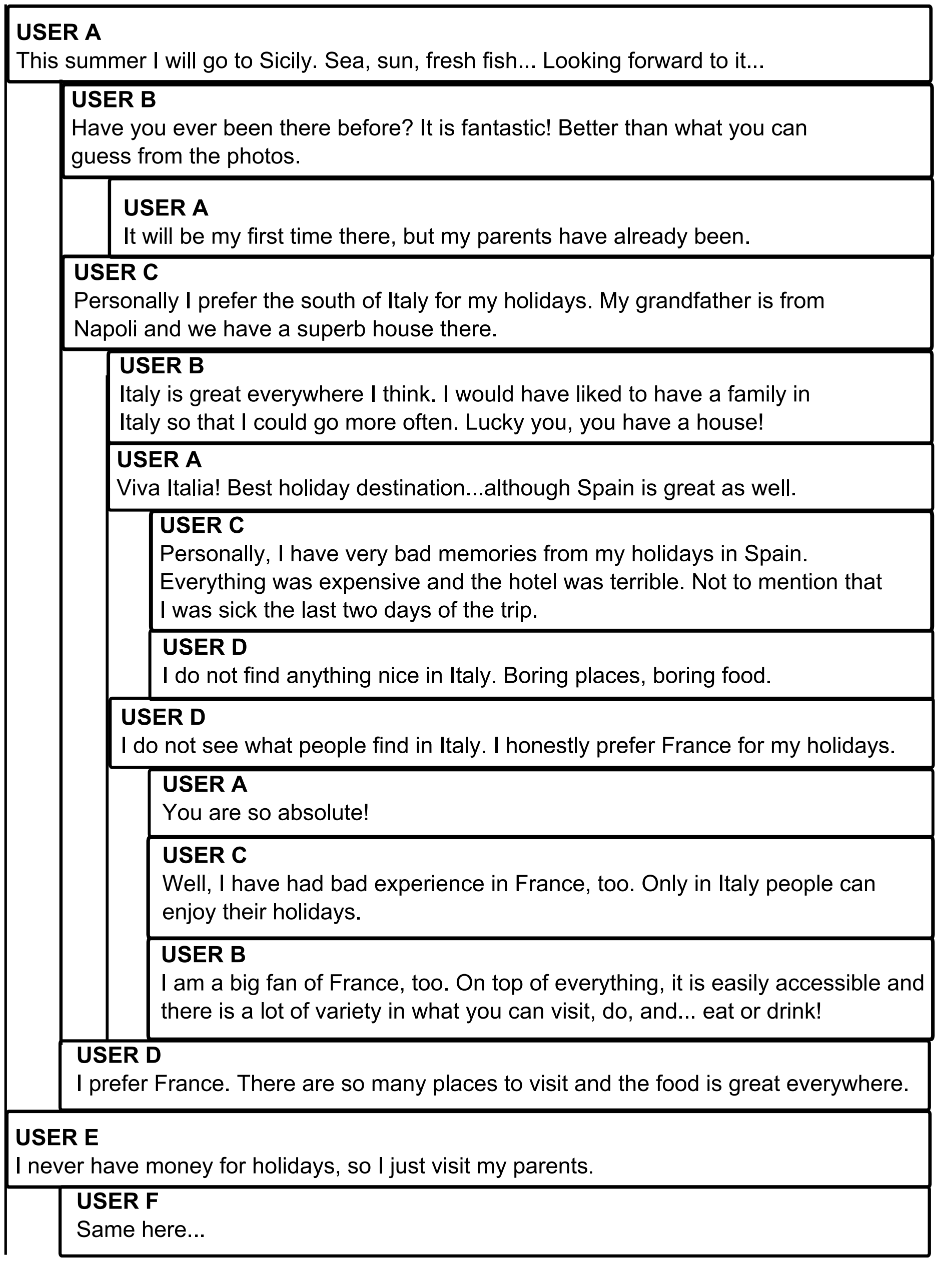


Figure 3.10: An artificial discussion.

Table 3.1: Message flow of the artificial discussion.

Posting	Author	Message	Reply-to	Opinion
v_1	A	“This summer...”	-	o
v_2	B	“Have you ever been ...”	v_1	p
v_3	C	“Personally I prefer...”	v_1	p
v_4	D	“I prefer France...”	v_1	p
v_5	A	“It will be my first...”	v_2	o
v_6	B	“Italy is great...”	v_3	p
v_7	A	“Viva Italia...”	v_3	p
v_8	C	“Personally, I have...”	v_7	n
v_9	D	“I do not find...”	v_7	n
v_{10}	D	“I do not see what...”	v_3	n
v_{11}	A	“You are so...”	v_{10}	n
v_{12}	C	“Well, I have...”	v_{10}	n
v_{13}	B	“I am a big fan...”	v_{10}	p
v_{14}	E	“I never have money ...”	-	o
v_{15}	F	“Same here...”	v_{15}	o

In Table 3.1, the column “Author” shows the author of the specific message, and the column “Message” shows the beginning of the actual message. The column “Reply-to” shows which posting the reply refers to and the column “Opinion” denotes whether there is a positive (p), negative (n), or neutral (o) opinion in the content of the message.

The representation of this short discussion by a PROG graph G is depicted in Figure 3.11. For comparison purposes the user-based graph is shown in Figure 3.12.

As we can see, the two graphs in Figure 3.11 and Figure 3.12 represent different information for the same discussion. The user-based graph shows the interaction between the discussion participants. We do not know who initiated the discussion or the order in which the users spoke to each other. Additionally we cannot identify the parts of the discussion during which the users have participated: did they speak only in the beginning or they were active participants throughout the whole discussion? For instance, in the user-based graph of Figure 3.12, we can see that the users A and B exchanged some messages but it is only in the PROG graph that we can identify during which part of the discussion they actually chatted. In the

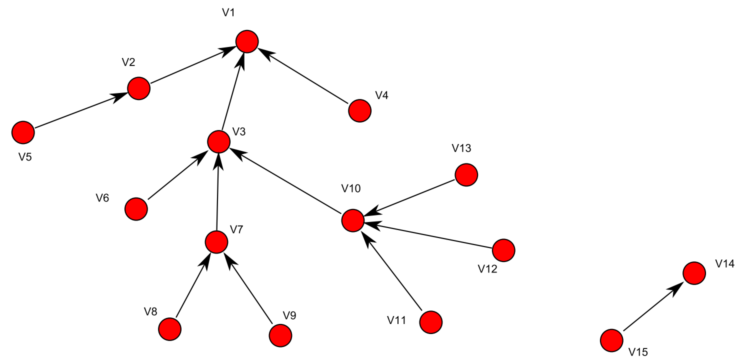


Figure 3.11: Post-Reply Opinion Graph of the artificial discussion.

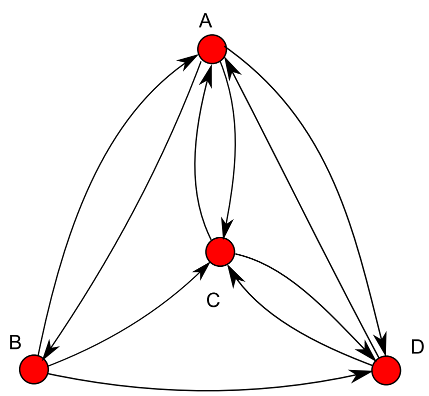


Figure 3.12: User-based graph of the artificial discussion.

PROG graph the messages are not just a bundle of random messages but they have a structure.

The PROG graph G of Figure 3.11 is consisted of two discussion threads ($orderThread(G) = 2$). We can distinguish the eight discussion chains of the biggest thread which are the following:
$\{v_1, v_2, v_5\}$, $\{v_1, v_3, v_6\}$, $\{v_1, v_3, v_7, v_8\}$, $\{v_1, v_3, v_7, v_9\}$, $\{v_1, v_3, v_{10}, v_{11}\}$, $\{v_1, v_3, v_{10}, v_{12}\}$, $\{v_1, v_3, v_{10}, v_{13}\}$, $\{v_1, v_4\}$.

The vertex which has started the discussion is the $initialVertex(G) = \{v_1\}$. As a result, we know that it is the author A the one who started the discussion. We can see in which parts of the discussion this author appears by looking at the $msgs('A')$ vertices. From this, we notice that the first discussion chain is just a short dialogue between the authors A and B since they do not speak to each other again in another discussion chain. This can be confirmed by looking at the content of the respective postings.

The graph allows us to notice the postings with the higher popularity. Let us consider the first type of popularity which is given by the *inDegree* value. The most popular messages found by this measure are the message objects 3 ($inDegree(v_3) = 3$) and 10 ($inDegree(v_{10}) = 3$). Both of these messages have caused reactions.

According to the second type of popularity that additionally takes into account the replies towards the replies of a posting, the most popular messages are the "root" one ($inDegreeExtra(v_1) = 12$) and then the posting v_3 ($inDegreeExtra(v_3) = 8$).

By the PROG representation we can identify the parts where opinion messages appear. In this example, all discussion chains contain some opinion information. More specifically, the vertex representing the posting v_{10} has received replies expressing both negative and positive opinions, and the vertex of the posting v_1 has only received reactions containing positive opinions.

In Table 3.2 we give the values of some measures per vertex. From this table, we identify which messages have caused reactions with positive or negative opinion polarities. The reactions of the posting v_{10}, for instance, are on average negative. Moreover, the average opinion values 1 (v_1) and -1 (v_7) show unanimous positive and negative opinion received respectively.

We can also see that the message objects that have received varied opinion replies have higher entropy than the rest of the message objects. In other words, the postings v_3 and v_{10} are regarded as the vertices of the graph that hold higher opinion information since they have caused varied opinion reactions.

The combination of the average message opinion value and the entropy reveals more information. For example, by knowing these two values for the popular postings v_3 and v_{10}, we can assume that the v_3 has received varied opinion reactions that are mostly positive, while the v_{10} has had various reactions mostly negative.

Table 3.2: Opinion measures applied to the artificial discussion.

v	$avgMsgOpinion(v)$	$reply(v,p)$	$reply(v,n)$	$reply(v,o)$	$H(v)$
1	1	3	0	0	0
2	0	0	0	1	0
3	0.33	2	1	0	0.28
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	-1	0	2	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	-0.33	1	2	0	0.28
11	0	0	0	0	0
12	0	0	0	0	0
13	0	0	0	0	0
14	0	0	0	1	0
15	0	0	0	0	0

In Table 3.3 we show the results of the opinion measures oriented towards the users. From this table, we notice that the user B has always had a positive reaction during the discussion. Also, the users C and D had a more negative than positive reaction. Furthermore, there was an average positive reaction towards the user A and C and an average negative reaction towards the user D. This is indeed the case in our example.

3.8 Analysis of a real online discussion

Applying our model to bigger discussions with many messages is very interesting in how it facilitates the mining of the discussion.

Table 3.3: Opinion measures applied to the users.

User u	$avgOpFromUsr(u)$	$avgOpToUsr(u)$
A	0	0.2
B	1	0
C	-0.33	0.33
D	-0.33	-0.33
E	0	0
F	0	0

In order to give a working example and evaluate in this way the proposed model and measures, we have taken a real web discussion from the site of a French newspaper (http://www.liberation.fr). The title of the discussion is “Pour ou contre les bureaux open space” (“For or against the open-space offices”). The discussion is in French and it consists of 120 messages. The users who have participated are 98. The actual message content of this discussion can be found in the Appendix B.

For this discussion, we have manually identified the opinion polarities since we have not found an available Opinion Mining tool that identifies opinion polarities in French text. We have automatically constructed the PROG graph which is shown in Figure 3.13. The image is taken from our prototype system which is described in Chapter 5. The vertices appear with an identification number calculated internally by our application. The light color (green) vertices show the vertices that belong to discussion threads with an order greater than 1. The darker color (red) vertices are the vertices that compose a discussion thread by themselves.

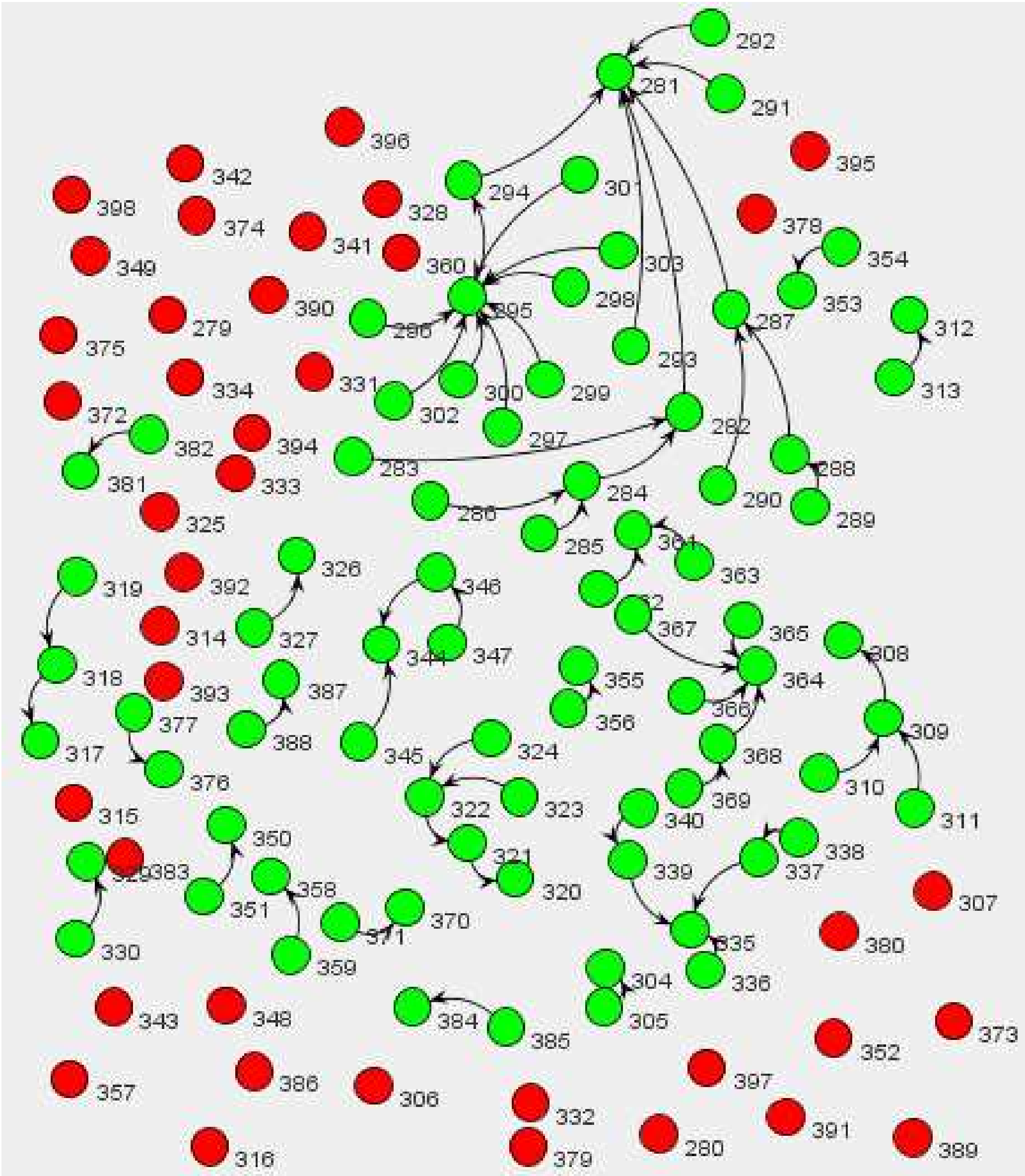


Figure 3.13: PROG graph of a real web discussion. The part with the light color vertices (green) consists of the discussion threads with an order greater than 1.

Let us see how we apply the defined measures and how they can help us to analyze the particular discussion.

3.8.1 Structure-oriented Measures

As we can see in Figure 3.13, the PROG graph is quite complex. We can discern a variety of vertices. The ones of the darker color (red) are "lonely" in the sense that they do not connect to any other vertex, and some others are involved in discussion threads.

The "lonely" vertices represent postings that do not reply to any other message and they have not received any reply either. These vertices are not very interesting for the discussion analysis since they have not played a crucial role for the development of the discussion. They are less probable to have an impact on the whole discussion or to contain interesting opinions. The "lonely" vertices are given by the set $root(G) \cap leaf(G)$, where both the root and the leaf measures refer to the vertices that initiate and terminate respectively a thread. In the particular discussion we have $\mid root(G) \cap leaf(G) \mid = 40$, so there are 40 vertices out of the 120 that we will put on the side, performing in this way a first step in the discussion mining.

In the particular discussion which is a forum, we have 21 discussion threads with $orderOfThread(G_{thr}) > 1$. Some of them are consisted of only two vertices. These could be small sub-discussions or dialogues between two users. They could be of question-answer type or a direct answer to an expressed statement. Examples of these 2-vertex discussion threads are between the messages (305, 304), (377, 376) or (327, 326) given here with the edge notation $e_{v'v}$.

Using the structure of the PROG graph, we can easily find out the messages that have started a discussion thread with more than one vertices. These are the following messages:

$root(G) - leaf(G) = \{281, 304, 308, 312, 317, 320, 326, 329, 335, 344, 350, 353, 355, 358, 361, 364, 370, 376, 381, 384, 387\}$.

These are 21 messages, a number that, of course, matches the number of discussion threads with a greater than 1 order. By looking at these messages we can get a quick summary of the main aspects discussed in the particular forum. We can observe just by reading the "root" messages that the discussion involves main advantages and disadvantages of working in an open-space office. The subjects vary between how a manager sees an open space as opposed to a simple employee (posting 329), a young person as opposed to a more experienced one (posting 376), aspects of the open space that may play a role such as the size, the other colleagues (postings 355, 353, 317, etc), a discussion about "anglicisme" that has involved many messages etc. We see also that the people who have participated come not only from France but also from other countries (posting 381). In conclusion, just from the $root(G) - leaf(G)$ messages we can discern what the people have discussed in the forum as well as the kind of people who have participated.

Let us, now, use the structure-oriented measures in order to "zoom" more into the discussion. By calculating the *inDegree* of the vertices of the graph G, we can identify the most popular messages. These messages together with their *inDegree* are given in Table 3.4 in descending degree order.

Table 3.4: The *inDegree* of the most popular messages.

v	$inDegree(v)$
295	8
281	6
364	4
335	3

By looking at the content of the most popular message, we see that it discusses about the distinction between the private and the professional life. This statement apparently has created a lot of discussion because it is not only the most popular message but it also belongs to the biggest discussion thread of the forum. The second most popular message (281) is the "root" of the same discussion thread. Then, we have a popular message about "anglicisme" and another one which shows a preference towards the offices that are not open-space.

3.8.2 Opinion-oriented Measures

The opinion presence inside the discussion interests us a lot during the discussion analysis. In Table 3.5 we show some general statistics of the discussion we analyze. From these statistics we notice that the general atmosphere in the discussion is negative, since there are more negative (n) opinion vertices than positive (p) ones.

Table 3.5: Statistics of the example discussion.

Postings	120
Users	98
Disc. Threads with order>1	21
Opinion Vertices	59 ($\mid n \mid > \mid p \mid$)

In Table 3.6 we show the results from applying some opinion measures on the most popular postings.

Table 3.6: Results of the opinion measures for the most popular messages.

v	$reply(v,p)$	$reply(v,n)$	$reply(v,o)$	$avgMsgOpinion(v)$	$H(v)$
295	1	6	1	-0.63	0.31
281	0	3	3	-0.5	0.3
364	0	0	4	0	0
335	0	2	1	-0.67	0.48

From Table 3.6, we see that the most popular message has had a total of 8 reactions, 6 of which were negative(n), 1 was positive(p) and 1 contained no opinion(o). The average opinion of three of the most popular messages is negative which is indicated also by the general tendency of the discussion. The entropy values which are not 0 show the existence of a variety in the opinion polarities of the reactions. Indeed, if we look at the content of the postings, this is the case.

In order to see the opinion of a user and towards a user, let us select the user u_p who has the most popular message of the discussion. This user has written $\mid msgs(u_p) \mid = 3$ messages in total in this discussion and his average opinion is $avgOpFromUsr(u) = -1$. This shows that whenever this user wrote a message he expressed a negative position. At the same time the replies towards this user have had an average $avgOpToUsr(u) = -0.67$ which shows that the position of the rest of the users that replied to this one was negative as well.

Similarly we can check the opinion status of the rest of the users.

3.8.3 Time-oriented Measures

In the specific forum that discusses about open-space offices, the very first message of the discussion is:

$$v = initialVertex(G) = 344$$

and it has been written on $tm_v =' 10/12/2008\ 19:19'$. The last message is:

$$v' = finalVertex(G) = 279$$

written on $tm_{v'} =' 29/12/2008\ 13:38'$.

From the first and the last message we can see the duration of the discussion. For the specific one the duration was:

$$duration(G) = 18\,days,\ 13\,hours\ and\,38\,minutes.$$

The duration of the discussion (almost 18 days) in comparison to the number of messages that have been posted (120) shows that the discussion has not been a very popular one.

The presence of the chronology in the graph facilitates the identification of the real "ancestors" of a posting, even if there is no actual link between them, connecting them in the graph. Let us use the $ancestors(v)$ measure in order to find out at which message the 355 posting replies to. Following the graph links, this message initiates a discussion thread and it replies to no other message. Inside its content, though, it seems that it replies to another message. By looking at its *ancestors* in combination with the knowledge of the author it replies to (extracted from the content of the message), we find out that the 355 message replies to the 374 posting.

3.8.4 Topic-oriented Measures

In order to find out the topics present in this discussion we have used the topic-extraction system AGAPE [VG07]. The algorithm has identified 5 topics present in this discussion. The topics together with the number of messages identified in them is shown in Table 3.7.

This is a small debate where each user has not posted a lot of messages. The maximum number of messages posted per user is 4 and 84 out of the 98 users have posted only a single message. Nevertheless we can calculate the average opinion per user and per topic.

Table 3.7: Topics identified by AGAPE [VG07] in the real discussion.

topic	Description	No. messages
1	space, personne, clavier, problème, point, raison, réalité, azerty, langue, machine, mot, type, plupart, question, titre	34
2	bureau, vie, chef, cas, entreprise, boite, besoin, fait, avantage, coup, intimité, manager, rapport, chose, internet, plateau, journée	29
3	temps, téléphone, collègue, travail, patron, spaces, bruit, jour, productivité, rêve, maison, condition, voisin, fonction, système, milieu, rien	24
4	gens, espace, an, anglicisme, part, avis, fil, evolution, 10heures, penser, famille, pain, soir, quotidien, mal,	11
5	monde, boulot, lieu, management, étude, fin, odeur, femme, configuration, panacée, cata, cactus, journaliste, doute, evidemment, début, toilette, café	11

In this discussion we have two users u_1 and u_2 who have more than one posting in topic 1, one user u_3 with more than one posts in topic 3 and one user u_4 with more than one posting in topic 5. Let us see their average opinion per topic:

$$\text{opEvolution}(u_1,\, 1) = 0.5$$
$$\text{opEvolution}(u_2,\, 1) = -0.5$$
$$\text{opEvolution}(u_3,\, 3) = -1$$
$$\text{opEvolution}(u_4,\, 5) = -0.5$$

As a result, we can identify that the user u_1 holds a more positive than negative opinion in topic 1, while the user u_3 holds a strong negative position in topic 5. In the same way we can see how the opinion of users changes from topic to topic.

3.8.5 User-oriented Measures

As mentioned before, the PROG graphs are not supposed to analyze the user interactions. Nevertheless, we can provide some user-oriented information using some of defined measures.

Looking again at the most popular messages, we can identify their authors. Then we can see whether these authors have other postings inside the discussion by the $msgs(u)$. In the specific discussion the user of the most popular message has sent another 2 postings in the discussion which have not created any particular reactions.

Another user-oriented measure is the identification of the people who are central for starting or for completing a discussion. The results are given by the $usersStartDisc(G)$ and the $usersEndDisc(G)$ respectively. In the discussion we are analyzing most users have participated only once, so these measures cannot give us a lot of information about the users.

3.9 Discussion

The proposed model and the various measures that it proposes for the purpose of a discussion analysis are evaluated through the application of this model to various online discussions. The evaluation is done by showing the advantages and the complementary information that can be extracted from a graph of type PROG as compared to the standard user-based graph of the social network model. One example of the model application to a real online discussion has been given in the previous Section. Through this example, we have seen how the proposed model differs from the social network model and what complementary information it provides. More online discussions have been evaluated through the prototype system we present in Chapter 5.

More specifically, from a graph of PROG type we can extract the following information:

The discussion chains and threads of the analyzed discussion. A graph of type PROG allows us to see how the discussion evolves and identify the threads and chains that form the discussion. The postings that appear in the same discussion chain or thread imply similarity in content. In a user-based graph that represents a social network, a "reply-to" relationship does not always mean similarity in topic since two users may have replied to each other many times in many different discussion chains or threads.

The popular postings which have caused many reactions. The PROG graph identifies the most popular messages. A message that has caused many reactions compared to one that has received none is definitely a more interesting message that may worth being analyzed in more detail. Social network graphs deal mainly with user-popularity and not postings-popularity.

The opinion polarity presence in the discussion. From the opinion identification point of view, the PROG graph facilitates the identification of the discussion parts that contain opinions, it enables the distinction between the objective and subjective sides of the discussion and it allows focusing on the parts that show more interest from an opinion exchange point of view. This permits the mining of the information and the focus of the analysis on a subset of the discussion rather than on the discussion as a whole. In addition, the new model, through the defined measures, permits measuring the average opinion during the discussion, the average opinion per discussion thread, as well as, the average opinion received per message. The sentiment behavior of a user and towards a user during the discussion can also be measured.

Moreover, having represented a discussion from the point of view of message objects instead of users allows us to identify quicker interesting postings. A posting that has received few but varied positive and negative opinions can be more interesting than one that has received plenty of replies that are all positive or neutral.

The indirect links between the postings by using the temporal and the topic information. Some postings inside a discussion are sent individually without showing a reply to an existing posting. Sometimes, these postings refer to new arguments or topics but some other times they intend to reply to previous postings or they have been influenced by them. By using the presented temporal and topic-based measures we can identify these previous postings that are related to the current one, even if the author of the particular posting has not directly shown it.

Main differences between the proposed and the existing representation are summarized in Table 3.8.

Table 3.8: Differences between a social network and the proposed Post-Reply Opinion Graph.

	Social Network	Post-Reply Opinion Graph
Entity	The main entity of the discussion is considered to be the *user* who participates.	The main entity is the *posting*.
Chains/Threads	-	Identify the postings that are connected in discussion chains or threads.
Interaction	We can observe how the users interact with each other.	We can observe how the message objects form discussion chains and threads.
Opinion	-	We can see the opinion flow of the discussion and measure the opinion information per posting, chain and discussion as a whole.
Popularity	If many edges arrive to a user vertex, then the specific user is popular because s/he has received messages by many people.	If many edges arrive to a PROG vertex, then the specific posting is popular because it has caused many reactions.

User-based and PROG graphs serve different purposes. Both of them give structure to a discussion and they aid the discussion analysis by extracting useful information from the structured representation. Using a combination of these graphs for the analysis of a discussion should be considered.

In conclusion, the Post-Reply Opinion Graph can be seen as a "zooming" process into the user-based one. The application of the new model to an online discussion results in the extraction of knowledge that cannot be provided or captured by the social network model. This shows the worth of the proposed model in the domain of discussion analysis.

3.10 Conclusions

In this Chapter, we have presented a new model for the representation of web discussions. The model is visualized through a Post-Reply Opinion Graph which enables a content-oriented analysis of a discussion. Using the definition of this graph, we have defined main concepts and measures which take advantage of the structure of the proposed graph as well as the information encapsulated in the model. We have defined structure-oriented, opinion-oriented, temporal-oriented and topic-oriented measures.

Every discussion that is found in the Web is particular and can be analyzed in different ways. A simple user may be interested only in the most popular messages or in the messages that contain opinion. A marketing analyst may be more interested in the messages that have caused some dispute or in those that belong to certain thematic categories. The measures can be chosen according to the task-in-hand and the role of the user who desires to browse through or analyze an online discussion.

Chapter 4

Recommendation of Useful Postings

In this Chapter, we propose using the information extracted from the Post-Reply Opinion Graphs in order to automatically classify the discussion postings according to how useful they can be to a user for a specific discussion. We experiment with various criteria and we focus on how they correlate with the way humans classify messages as interesting and non-interesting. Such a classification allows recommending to a user the more interesting messages among all the rest. In this way, a user can quickly get an idea of the content of the discussion and identify how to participate or to whom to talk to in the first place. The experiments carried out with real forums found on the Web in both French and English, have shown that the proposed framework combined with the right criteria allows the identification of interesting messages.

4.1 Introduction

The Post-Reply Opinion Graph allows us to define criteria so as to determine the most interesting or key messages that appear inside an online discussion. The recommendation to an end-user of a list of the most interesting discussion messages would help the user navigate quicker and more efficiently inside the discussion. In this Chapter, we consider the user to be one that wants to get an idea of the content of the discussion and participate as well, if possible.

Key is considered to be a message that has an impact on the whole discussion. We assume that human preferences are correlated, otherwise recommendations and predictions could not be possible [PHLG00]. Most end-users assume a message to be interesting if it follows at least one of the following assumptions[1], which are presented here in descending order of popularity:

Opinion: A message is interesting when it contains opinion.

Size: A message is interesting when it is present inside a long discussion thread.

Reactions: An interesting message has caused many reactions.

Initial: The initial message of a discussion thread is interesting.

Time: The most recent message is interesting.

The distinction between the various messages permits a kind of summarization of the discussion and it allows the user to get an idea of the content and the main ideas that have been expressed. Considering the fact that most important discussions can be very long, with hundreds and thousands of messages, our model helps the user to identify quickly how to participate or to whom to start talking to, since there is no need to spend time reading the whole discussion.

Our main objective is to distinguish between the interesting and the non-interesting messages that are posted in an online discussion. Let us consider M to be the set of the discussion messages. We are looking for a set of messages M' to be recommended to the end-user such that the following conditions are satisfied:

- $M' \subset M$, so only a subset of the whole set M can be proposed to the user and
- the set M' should contain the messages which the user would select as the most useful.

[1]The assumptions were collected after having interviewed 10 end-users who visit online discussions such as forums in an almost daily basis.

The need to identify and extract key messages from discussions with the purpose of recommending them to end-users leads us to study the field of recommender systems. The recommender systems are systems which use the knowledge they have about user profiles in order to recommend items to users such as products, documents, news articles. In our case the recommendation concerns discussion messages.

The contributions of this Chapter are summarized in the following:

1. We use the Post-Reply Opinion Graph in order to extract key messages from online discussions.

2. A number of criteria are studied and they are correlated with user preferences.

3. Extensive experiments are carried out which allows us to see the conditions under which a recommendation set is acceptable by a user.

In this Chapter, we begin by giving a brief overview of recommender systems in order to point out their main characteristics. Then, in Section 4.3 we study how we can use the concepts of recommender systems for our benefit. Section 4.4 deals with the definition of criteria based on the Post-Reply Opinion Graph model and in Section 4.5 we experiment by extracting key messages from real forums using these defined criteria. Section 4.6 concludes.

4.2 Related Work

This Section is a quick overview of the main characteristics of recommender systems. We present examples of recommender systems whose task relates to ours in the sense of distinguishing between interesting and non-interesting items (documents, e-mails, news).

The recommender systems are systems whose main goal is to recommend items to users. These items can range from movies and news articles to proposals for holiday destinations or for meeting people. In such systems, the users can usually rate the items according to how much they prefer them. Each user is treated differently according to his/her profile and preferences.

According to [AT05] the recommendation problem is formally defined as following:

Recommendation problem. Let us consider a set of items I and a set of persons P. We want to choose an item $i' \in I$ for the person $p \in P$ such that the utility u of the person is the maximum:

$$i'_p = argmax_{i \in I} u(i, p).$$

The utility can be a rating scheme or any profit function.

In our case which is recommending interesting messages extracted from discussions, the utility function represents the satisfaction of the user regarding how interesting a message is.

We have three types of recommender systems [AT05]:

- *Content-based*: These recommender systems base their predictions on the similarity between the items already rated by the user. The content of the items is often described by weighted keywords and the user-profile is seen as a vector of weights, where each weight represents the importance of a respective word to the user.
- *Collaborative*: They base their predictions on the similarity between the users that have already rated certain items. The similarity can be measured by correlation metrics (e.g. Pearson), by identifying the nearest neighbors of a user or by other similarity measures. Models can also be learnt in order to make rating predictions.
- *Hybrid*: These are recommender systems whose techniques use a combination of both the content-based and the collaborative methodologies.

One example of a content-based recommender system is the Syskill & Webert system presented in [PB97]. The purpose of this system is to recommend interesting web pages on a particular topic. The system learns user profiles that differ per topic. The profile of a new user is constructed by asking the user to provide some words that would characterize interesting-on-the-topic pages and with what probability. The system identifies the most informative words of each web page by calculating the expected information gain that the presence or absence of a word gives towards the classification of elements of a set of pages. Then, it applies a naïve Bayes classifier in order to learn interesting pages. The user can rate the recommended pages as positive or negative examples and these examples are used when a profile is learnt.

Among the first collaborative recommender systems is Tapestry [GNOT92]. This is a system in which a filtering of e-mails and news takes place so that only interesting to the user items are kept. The system relies on direct as well as implicit user feedback. Implicit feedback is based on user actions. For example, a mail that is replied to by an interesting user is an interesting mail since it has been given attention to by this user. Tapestry is filtering items into two steps; the first step includes separating the items in “good” or “bad” and the second step is prioritizing the “good” items. GroupLens [RIS+94] is another early collaborative recommender system which pointed out that studying the user profiles and the ratings they give to certain items can improve the ability of a system to recommend interesting news articles to users.

Fab [BS97] is an example of a system that combines both content-based and collaborative techniques and it is made to recommend web pages to users. In this system, pages are collected by topic and then they are forwarded to users according to their profiles. The users provide ratings from a 7-point scale for the pages they have already seen and these ratings are stored into the system in order to be used for future recommendations based on similarity of users or items.

Another work which is outside the recommender systems domain but it is highly related with ours, is a quality document identification system [Elk07]. This system classifies documents to important and less important. It classifies mainly text documents with an application of ranking messages in discussion groups according to quality. The classifier is based on various criteria such as the vocabulary used, the length of words/sentences and the usage of grammar. The training data are collected with the help of humans who determine which types of documents/messages are of high quality. This work, though very interesting, differs from ours in that we do not use a learning system. We are based on the structure of the discussion and its graph representation rather than on the low-level features of a text (e.g. average length of words, usage of punctuation, orthography etc.).

4.3 The “cold-start” case and our approach

In this thesis, the fact that the users have different preferences and provide ratings to items is not taken into account. No actual user profiles are logged. Our approach is based on the structure of the discussion and the assumptions regarding what a key message is that are previously presented. As a result, we cannot consider our approach to be that of a proper recommender system.

Nevertheless, there are a lot of similarities with the needs of the recommender systems and, therefore, we can define our problem within the domain of recommender systems.

In a recommender system there are two basic concepts; the *user-profiles* and the *item-profiles*. According to [AT05], other dimensions need to be used in order to accurately recommend items, such as the temporal dimension and the knowledge about the user's task i.e. why the user has logged into the system.

In our case, the user-profiles contain just a user-ID, but in the future they can be populated with user characteristics such as the age, the gender, the education status. Although, we have not studied, in the scope of this thesis, how the age or similar characteristics could affect the choice of interesting messages, we have observed that people of similar age and education status tend to consider the same messages as interesting.

The items are actually the different discussion messages. The item-profiles contain an id, an author, the message itself and the time the message was written.

Choosing the criteria for providing key messages to users, can be considered as the "cold-start" case of a recommender system. This is the case when a new user enters the system and the system knows nothing about this user. It does not know any of the user's likes/dislikes and as a result, it cannot match the user with already known user profiles. In such cases, a system has to deal with very little data.

In the case of selecting messages from a discussion, the system can use information from the structure of the discussion such as item popularity (inherently a message that has been replied to is considered to be interesting), or the entropy (how much information is included in the message). The entropy has already been used for new-user-cases ([KM01]) in order to construct a list of items with the purpose of supplying it to the user for rating. The entropy helped in finding out which item would give the maximum information if it was rated by the user, so that this item is included in the list. The entropy has also been used in [RAC+02] in combination with the popularity in order to identify items that may interest the users of the system.

In conclusion, the approach we want to evaluate is actually that of a recommender system which lacks user profiles.

4.4 Selection Criteria

Using the structure and the information encapsulated in a Post-Reply Opinion Graph, we can extract useful knowledge regarding the discussion. For the purpose of differentiating between the degrees of importance for each message, we have explored a series of criteria. These criteria have been applied separately to sets of messages.

Taking into account the assumptions we have previously made which classify a message as interesting or not, we consider the significance of a discussion posting to be a function of:

- the number of vertices that exist in the thread it is part of i.e. the order of the discussion thread,
- whether it initiates a thread or not,
- its popularity i.e. the number of responses it has received,
- the opinion it holds,
- whether the reactions it has caused contain opinion and
- the variety in the opinion of the reactions it has caused.

The algorithm we have developed for the selection of key messages receives as input the message postings of a discussion, where the "reply-to" links and the opinion polarities of the postings are known and the criteria we want to apply. The output of the algorithm ranks all postings according to whether they satisfy the selected criterion. For example, if the criterion of popularity is applied, the postings with a higher popularity are higher in the ranking list than the others. Postings with the same criterion value are sorted in descending order of posted time i.e. the most recent message goes higher in the rank. In order to extract the most interesting messages, a threshold can be applied to the algorithm so that it retrieves a maximum number of messages.

In the following sections we present the criteria used in order to extract key messages from a discussion which is represented by a Post-Reply Opinion Graph.

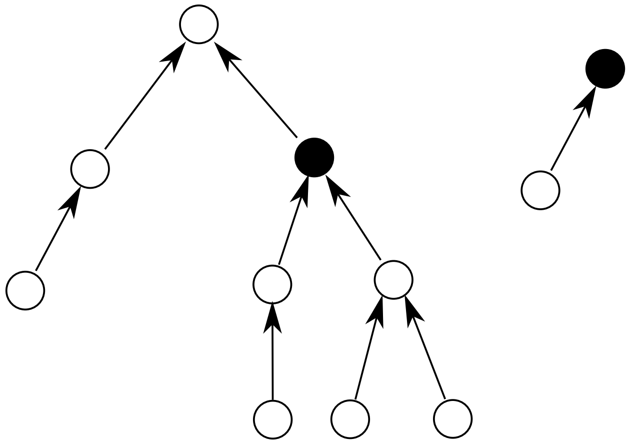


Figure 4.1: The importance of a vertex differs according to the order of the discussion thread it belongs to. The black vertex on the left hand side has played a more significant role in the discussion than the black vertex on the right hand side.

4.4.1 Order of a discussion thread

A posting that belongs to a discussion thread with many vertices does not have the same importance as a posting that belongs to a thread with one or two vertices. In Figure 4.1, for example, the black vertex on the left-hand side of the Figure may have influenced a part of the discussion, while the black vertex on the right-hand side of the Figure does not have the same weight.

Each discussion thread consists of a number of vertices. The order of the discussion thread G_{thr} where the vertex v belongs to is defined as:

$$orderThr(G_v) = \mid V_{thr} \mid,$$

where $G_v = (V_{thr}, E_{thr})$ is the subgraph of G that represents the discussion thread G_{thr} which contains the vertex v.

We define a criterion that retrieves a set of messages which belong to a thread with a minimum order, defined by the threshold d. A vertex v of a graph $G = (V, E)$ will be part of the set of these messages if it is retrieved by the following criterion:

$$order(G, d) = \{v \in V : orderThr(G_v) > d\}, \tag{4.1}$$

where G_v is the discussion thread of the graph G that contains the vertex v, and d is a threshold that can be either calculated automatically or defined by the user.

In the rest of the paper we refer to this criterion as "order".

4.4.2 Root Vertices

A discussion is often divided into sub-topics, since discussion participants tend to discuss about a particular topic or elaborate on a specific argument. A user that wants to speak about an argument/topic that has not been referred to until then, may initialize a new discussion thread by sending a new message that is not a reply to an existing post. This would be a "root" message defined in the previous Chapter by the $root(G)$ function.

"Root" is also a message that has never received any reply, and as a result it may not have influenced the flow of the discussion. Such a message may have a different degree of importance from a message which is the root of a long discussion thread. If we want to differentiate between "lonely" root vertices (i.e. vertices which belong to a discussion thread of order 1) and the rest of the root vertices, we can define the criterion as following:

$$root(G, d) = \{v \in V : outDegree(v) = 0, orderThr(G_v) > d\}, \tag{4.2}$$

where G_v is the discussion thread that contains the vertex v and d is a user-defined or automatically calculated threshold that points out the minimum desired number of vertices of the discussion thread.

In the rest of the paper we refer to this criterion as "root".

4.4.3 Vertex Popularity

As mentioned in the previous Chapter, the *inDegree* of a vertex captures its popularity. The higher the popularity, the more attention the message represented by the vertex has been paid to.

A vertex v of a graph $G = (V, E)$ can be considered to be popular if it satisfies the following criterion:

$$popular(G, d) = \{v \in V : inDegree(v) > d\}, \tag{4.3}$$

where $d > 0$ is a threshold that can be either calculated automatically or defined by the user.

A post that has received 10 replies does not have the same importance for the flow of a discussion as a post that has received none. Of course, the position of the vertex in the discussion thread plays a role as well [SVC09].

This criterion will be referred to as "popularity".

4.4.4 Opinion Content

One advantage of a Post-Reply Opinion Graph is that it encapsulates the opinion information of each post. A user that scans a discussion with many posts is looking for messages that contain opinion or messages where there is a presence of arguments. As a result, a message that contains opinion is more probable to be a key message than a message that is just informative (subjective message).

The opinion messages are represented by vertices $v = (m_v,\ op_v,\ u_v,\ tm_v)$, where $op_v \neq o$.

This criterion will be referred to as simply "opinion".

4.4.5 Opinion Reactions

Another criterion we define that uses the opinion information of a Post-Reply Opinion Graph is the number of reactions which contain opinion.

The $\sum_{r=n,p}(reply(v, r))$ is the number of replies that hold an opinion. The $reply(v, p)$ and $reply(v, n)$ have been defined in the previous Chapter and they denote the number of replies expressing a positive and a negative opinion respectively.

This criterion is an indication of whether a post has caused reactions that contain opinion or just information and it is 0 only if all reactions are opinion-free.

A vertex v of a graph $G = (V, E)$ can be considered to satisfy the criterion of opinion reactions if it belongs to the set:

$$reply(G, d) = \{v \in V : \sum_{r=n,p}(reply(v, r)) > d\}, \tag{4.4}$$

where $d > 0$ is a threshold that can be either calculated automatically or defined by the user.

The criterion will be referred to as "reply".

4.4.6 Entropy

This criterion is similar to the previous one that counts the reactions with opinion, but it is not a counter. The entropy H of a node v measures the variety in the opinion of the replies a message has received and it has been defined in the previous Chapter as:

$$H(v) = -\sum_{r=n,o,p}\left(\frac{reply(v,r)}{inDegree(v)}\log\frac{reply(v,r)}{inDegree(v)}\right).$$

According to this definition, if a vertex has received a number of replies that are all of the same opinion orientation, then the entropy will be 0. This criterion captures phenomena where a particular posting has caused disagreement or replies with various opinion degrees of arguments. Such a posting can be interesting for a discussion analyst in order to investigate the reasons why people argue.

A vertex v of a graph $G = (V, E)$ can be considered to satisfy the criterion of entropy if it belons to the set:

$$entropy(G, d) = \{v \in V : H(v) > d\}, \tag{4.5}$$

where $d > 0$ is a threshold that can be either calculated automatically or defined by the user.

This criterion will be referred to as "entropy".

4.5 Evaluation

The evaluation of a system that extracts key messages from discussions is not straightforward. Generally, the evaluation of recommender systems is a difficult issue. The reasons are mainly a)the "variety of datasets" (variety in number of users, items, ratings) and b)the "error limit" which points out that there is a limit on how good the results we have may be, since the same person may give different ratings for the same item in different times [HKLJ04, SM95].

In the case of online discussions, these issues are interpreted as following:

a) *variety of datasets*: each discussion is different. There is a variety in the number of users, the number of messages, the distribution of messages and users, the content of the discussion (sometimes it is more opinion-oriented than others), the style of the language used etc.

b) *error barrier*: the users can change their mind as to which message is interesting, they may be affected by which messages they read first, how focused they are etc. These characteristics are reflected on the evaluation in the sense that the results cannot surpass a certain limit.

One of the most important decisions before evaluating a system is to identify what we are expecting from it. In our case our goals are summarized in the following:

- Identification of interesting messages among the discussion messages.
- Rejection of the messages with no interest/quality.
- Recommendation to the users of a set of messages that could help them navigate inside the discussion.

An evaluation could be done in two ways; an implicit and an explicit one. The implicit way could involve monitoring of the behavior of users when they are presented with the recommended messages. If they actually click on the proposed messages, this could be an initial indication that they find these messages interesting. If they click on them, then we could monitor how much time they do spend on them, or whether they actually follow the specific discussion chain in order to get more information about what has been said. The implicit methods of evaluating a system are not reported to be accurate. This issue in combination with the fact that our system is not yet put to production so as to have many users that interact with it, does not allow us to present such an evaluation here.

The evaluation we perform is an explicit one which means that we directly ask the experts to rate the interesting messages. We do it in two parts:

1. We initially ask the experts to label the messages that they consider to be key and then we compare the labeled messages with the messages the system found as key.
2. We show the messages identified as key by our system to the experts and the experts tell us whether they agree or not.

The first part of the evaluation helps us to analyze how the different criteria or a combination of criteria can be applied to real online discussions in order to help us distinguish and extract interesting messages. In addition, it helps us identifying whether these criteria correlate with how the humans proceed in classifying the messages as key ones or not. The second part of the evaluation confirms that the criteria used are actually approved by the users when it comes to selecting key messages.

4.5.1 Experimental Dataset

The initial experiments that have been carried out involve French forums from the http://www.liberation.fr/forums/ web site. This site contains discussions about politics, economics and the society in general. In these discussions the users identify themselves by user names. The "reply-to" links between the messages are known.

We have analyzed a total of 1,147 messages that appear in eight forums. Some statistics about the forums are shown in Table 4.1. From this Table, we can observe that on a total of 636 replies, only 187 contain opinions.

Table 4.1: The experimental set of forums.

Total messages	1,147
Total discussion threads	510
Total reactions	636
Total opinion reactions	187

The messages were manually annotated with opinion polarities, since the automatic Opinion Mining is considered out of scope for the particular evaluation. We applied each of the pre-mentioned criteria to the set of the messages of each forum, and we identified the key messages per forum and per criterion.

In order to evaluate the results given by each criterion, we asked human raters to classify the messages as being key or not for the flow of the discussion. For each forum, two experts identified the key messages. The experts were free to choose an unlimited number of key messages per discussion. In Table 4.2 we see the number of messages selected as key per expert and per forum as well as the total number of messages of the forum. It has to be noted that the person mentioned as "Expert 1" was not always the same person. The same stands for "Expert 2". This means that we had two evaluations per forum but the experts were not always the same people.

Table 4.2: Number of messages per forum and number of key messages selected per expert.

Forum	Total Messages	Expert 1	Expert 2
1	91	17	7
2	120	18	8
3	274	6	15
4	45	5	7
5	272	14	28
6	66	3	17
7	245	22	40
8	34	1	9

The Pearson correlation between the experts is shown in Table 4.3. As a result of knowing the age and status of the experts that did the classification, we noticed that in cases where the two experts are of similar age and status they tend to achieve higher correlation. This is related to personalization issues ([EV05], [Mob07]) that should be investigated in the future.

Table 4.3: Correlation between 2 human raters per forum.

Forum	1	2	3	4	5	6	7	8
Correlation	0.28	0.45	0.40	0.43	0.09	0.20	0.25	0.29

4.5.2 Evaluation Metrics

Our task is like the "annotation in context" task mentioned in [HKLJ04]. This task involves identifying among structured discussion postings the ones that worth being read. For such a task, the most appropriate metrics to be used are the *classification accuracy metrics* [HKLJ04] which measure how often does the system decide correctly.

The metrics that belong to this category are the common *precision*, *recall* and *F1-measure*.

The *precision* shows how many of the messages found by our system were actually chosen as key by the expert. In other words, it shows whether we avoided retrieving non-key messages and it is given by the formula:

$$precision = \frac{correctly\ assigned\ key\ messages}{total\ key\ messages\ found\ by\ the\ system}.$$

The *recall* is a measure that represents the coverage of our algorithm. It shows how many of the messages that were considered interesting by the experts, were also retrieved by our criteria and it is given by:

$$recall = \frac{correctly\ assigned\ key\ messages}{total\ key\ messages\ found\ by\ the\ user}.$$

The *F1-measure* used in this Chapter is the one given by the formula

$$F1 = \frac{2*Recall*Precision}{Recall+Precision}.$$

4.5.3 Choosing the right threshold

One of the decisions that have to be taken when evaluating each criterion is the threshold d to be used. The threshold makes sense for the criteria where there is some granularity. For criteria such as the "opinion", where the answer is binary (0 or 1), all thresholds give the same results.

Let us consider a particular discussion represented by the graph $G = (V, E)$, for which the maximum value of a criterion C is M, and the value for the respective criterion for a vertex v is $criterion_C(v)$. Let also the threshold be t.

Definition. The posting represented by the vertex $v \in V$ can be added in the list of recommended postings if, for the criterion C and a predefined threshold t the following holds:

$$\frac{criterion_C(v)}{M} \geq t$$

Let us assume that for the criterion "popularity", we have set the threshold to be 0.2. A threshold of 0.2 means that for a discussion where the maximum criterion value (in this case for the criterion "popularity") is 6, we only put in the recommended list the vertices (postings) whose minimum value for the criterion is equal to 2. For the criterion of popularity, we would retrieve the postings with minimum 2 replies (since $1/6 < 0.2$ but $2/6 > 0.2$).

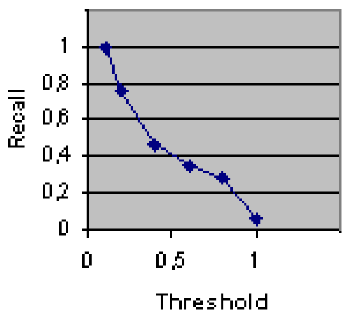
We experimented with various thresholds $t \in (0, 1]$, such as 0.1, 0.2, 0.4, 0.6, 0.8 and 1, in order to study the difference in our results. For each criterion we tried to choose the threshold that minimizes the number of messages retrieved while giving good performance results. A threshold of 0.1 results in the extraction of more messages than the usage of a threshold equal to 1. Extracting more messages gives a better recall, but it could be frustrating for the end-user because it would lead to the recommendation of too many messages. On the other hand, a threshold equal to 1, retrieves the vertices that have only the maximum value for the particular criterion and although it results in the extraction of few messages, the recall is very low.

An example that shows the differences in results when setting different values to the threshold is presented in Table 4.4 for a particular forum of 91 messages and for the criterion "order".

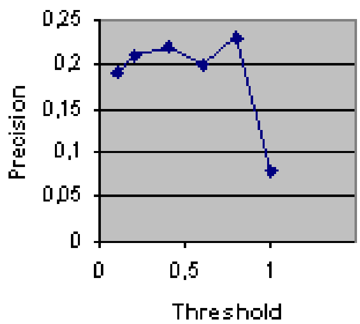
Table 4.4: Recall, Precision and F1-measure results per threshold for the criterion "order" applied to a forum with 91 messages.

Threshold	Messages	Rec	Prec	F1
0.1	91	1	0.19	0.32
0.2	62	0.76	0.21	0.33
0.4	36	0.47	0.22	0.30
0.6	30	0.35	0.2	0.25
0.8	22	0.29	0.23	0.26
1.0	12	0.06	0.08	0.07

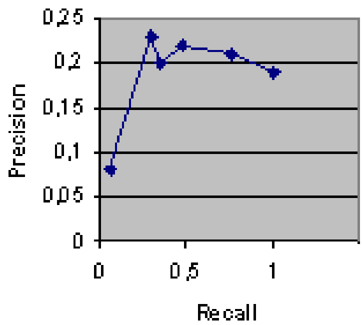
In Table 4.4 the column "Messages" shows the number of messages that are extracted as key when the respective threshold is set. We can observe that the higher the threshold the less messages are retrieved. This has an impact on recall as well as precision. The higher the threshold, the lower the recall, since the number of postings retrieved by the application of the particular criterion reduces. At the same time, as the recall decreases, the precision increases but only until a certain point that signals the limitation of the messages that can be retrieved. We present the threshold-recall, threshold-precision, and the recall-precision charts (Figure 4.2) for the results of Table 4.4. From these results we notice that the precision is quite low for the particular criterion. Later on, we will see how these results improve when we perform an aggregation of the criteria.



(a)



(b)



(c)

Figure 4.2: Threshold-Recall, Threshold-Precision and, Recall-Precision plots for the criterion "order" applied to a forum with 91 messages.

Threshold for the criterion Order

The plot that shows the average F1-measure values per threshold for the criterion "order" is shown in Figure 4.3.

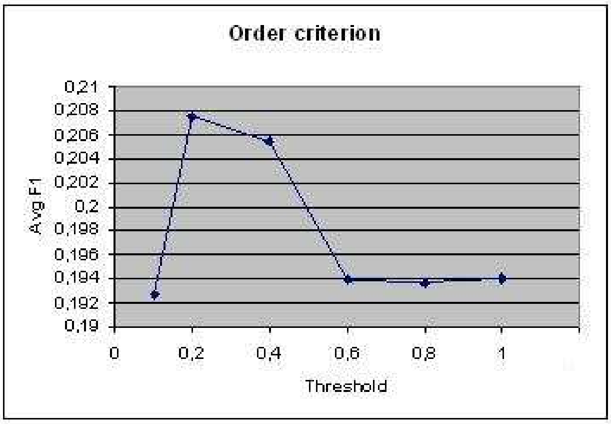


Figure 4.3: The average F1-measure per threshold for the criterion "Order".

From Figure 4.3, we can see that the maximum value for the F1-measure is achieved for a 0.2 threshold.

The issue with such a low threshold is that in some cases we may have the retrieval of too many messages. As a result, after having applied this criterion, we need to reduce the number of messages that can be recommended to a user. This is achieved later on with the aggregation of criteria.

Threshold for the criterion Root

We choose not to apply a threshold to this criterion so that even the nodes which result to discussion threads with an order equal to 1 are considered. As a result the "root" criterion is independent of the threshold.

Threshold for the criterion Popularity

The plot that shows the average F1-measure per threshold is shown in Figure 4.4. From this Figure, we can see that the maximum value for the F1-measure is achieved for a 0.2 threshold. Similarly to the "Order" criterion, with such a low threshold we may have the retrieval of too many messages.

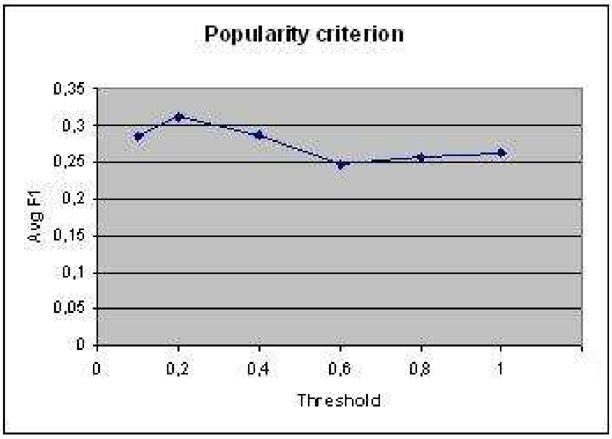


Figure 4.4: The average F1-measure per threshold for the criterion "Popularity".

Threshold for the criterion Opinion

Independent of the threshold.

Threshold for the criterion Opinion Reactions (Reply)

The plot that shows the average F1-measure per threshold is shown in Figure 4.5.

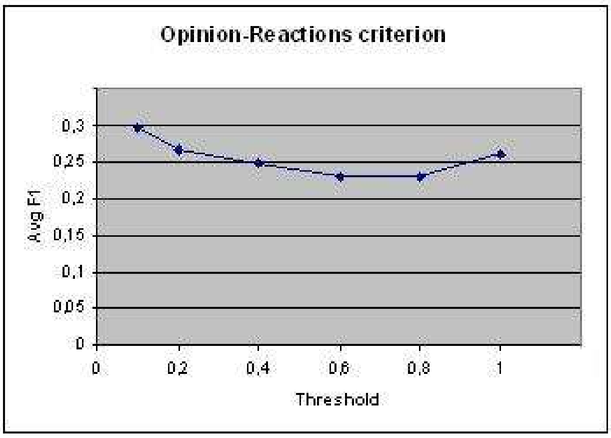


Figure 4.5: The average F1-measure per threshold for the criterion "Reply".

From Figure 4.5, we can see that the maximum value for the F1-measure is achieved for a 0.1 threshold. Although this threshold results in the extraction of many messages, it is leveraged by the fact that this criterion is hard to be satisfied and, thus, there are not many messages that comply with it.

Threshold for the criterion Entropy

The plot that shows the average F1-measure per threshold is shown in Figure4.6.

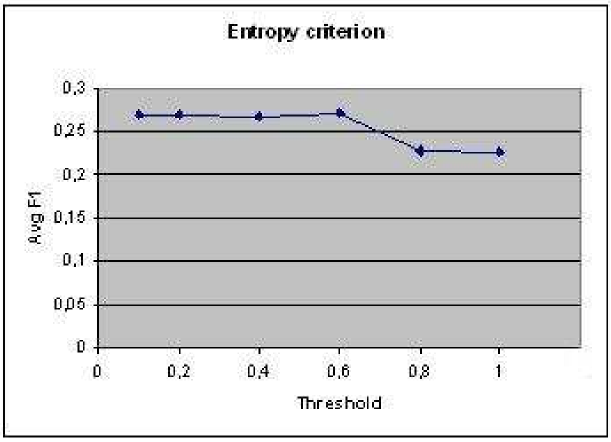


Figure 4.6: The average F1-measure per threshold for the criterion “Entropy”.

From Figure 4.6, we can see that the maximum value for the F1-measure is achieved for a 0.6 threshold.

4.5.4 Evaluation of each criterion separately

We looked at the standard recall, precision and F1 measures in order to evaluate the correlation between the results of each criterion and those of the human raters. For each criterion we used the thresholds mentioned in the previous Section. Therefore, we have chosen a 0.2 threshold for the “order” criterion, a 0.2 threshold for the “popularity” criterion, a 0.1 for the “reply” criterion and a 0.6 threshold for the “entropy” criterion. As mentioned previously, the rest of the criteria do not have a granularity and as a result the threshold usage has no sense.

The results of recall, precision and F1-measure per forum, per user and per criterion are presented in Table 4.5 and Table 4.6.

Table 4.5: Recall, Precision and F1-measure results per criterion (**Order, Root, Popularity**) when the optimum threshold value per criterion is used.

Forum	Expert	Order			Root			Popularity		
		Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1
1	1	0.86	0.10	0.28	0.86	0.12	0.21	0.86	0.18	0.28
	2	0.76	0.36	0.49	0.82	0.28	0.42	0.71	0.36	0.48
2	1	0.50	0.10	0.17	0.50	0.07	0.12	0.13	0.09	0.11
	2	0.50	0.23	0.32	0.44	0.13	0.20	0.17	0.27	0.21
3	1	0.67	0.07	0.13	0.73	0.07	0.13	0.33	0.33	0.33
	2	1.00	0.06	0.11	0.83	0.05	0.09	0.83	0.33	0.47
4	1	0.60	0.09	0.16	0.60	0.25	0.35	0.30	0.33	0.31
	2	1.00	0.20	0.33	0.14	0.08	0.10	0.29	0.33	0.31
5	1	0.36	0.06	0.10	0.71	0.07	0.13	0.07	0.04	0.05
	2	0.32	0.12	0.17	0.68	0.13	0.22	0.29	0.31	0.30
6	1	1.00	0.05	0.10	0.67	0.12	0.20	1.00	0.30	0.46
	2	0.82	0.23	0.36	0.71	0.71	0.71	0.29	0.50	0.37
7	1	0.73	0.10	0.18	0.50	0.22	0.31	0.18	0.19	0.18
	2	0.50	0.13	0.21	0.48	0.39	0.43	0.10	0.19	0.13
8	1	1.00	0.04	0.08	1.00	0.07	0.13	1.00	0.09	0.17
	2	0.67	0.25	0.36	0.56	0.38	0.45	0.56	0.45	0.50
Average		**0.71**	**0.14**	**0.22**	**0.64**	**0.20**	**0.26**	**0.44**	**0.27**	**0.29**

Table 4.6: Recall, Precision and F1-measures results per criterion (**Opinion, Reply, Entropy**) when the optimum threshold value per criterion is used.

Forum	Expert	Opinion			Reply			Entropy		
		Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1
1	1	0.86	0.14	0.24	0.57	0.36	0.44	0.14	1.00	0.25
	2	0.76	0.31	0.44	0.35	0.55	0.43	0.06	1.00	0.11
2	1	0.86	0.12	0.21	0.13	0.06	0.08	0.13	0.17	0.14
	2	0.72	0.22	0.34	0.28	0.31	0.29	0.17	0.50	0.25
3	1	0.53	0.09	0.15	0.27	0.22	0.24	0.13	0.29	0.18
	2	1.00	0.07	0.13	0.33	0.11	0.17	0.33	0.29	0.31
4	1	0.60	0.21	0.31	0.20	1.00	0.40	0.40	1.00	0.57
	2	0.71	0.36	0.48	0.43	0.60	0.50	0.29	1.00	0.44
5	1	0.21	0.04	0.07	0.21	0.11	0.15	0.07	0.09	0.08
	2	0.25	0.10	0.14	0.21	0.22	0.22	0.11	0.27	0.15
6	1	0.33	0.06	0.10	0.67	0.18	0.29	0.67	0.50	0.57
	2	0.47	0.47	0.47	0.29	0.45	0.36	0.12	0.50	0.19
7	1	0.59	0.17	0.26	0.55	0.32	0.40	0.32	0.41	0.36
	2	0.40	0.21	0.28	0.28	0.29	0.28	0.18	0.41	0.25
8	1	1.00	0.50	0.67	0	0	-	0	0	-
	2	0.11	0.5	0.18	0.11	1.00	0.20	0.11	1.00	0.20
Average		**0.59**	**0.22**	**0.28**	**0.31**	**0.36**	**0.30**	**0.20**	**0.53**	**0.27**

In Table 4.5 and Table 4.6, columns 3 and 4 give the recall and the precision when the only criterion applied to the list of forum messages is the order of the discussion thread. The results indicate that this criterion achieves the highest average recall and the lowest average precision of all the criteria. This can be attributed to the fact that important messages tend to be involved in long discussion threads, since they cause reactions and sub-discussions. The low precision, though, is because not all messages of long threads are considered to be key. Similarly, the results of columns 6 and 7 show that when the root criterion is applied, the recall is much higher than the precision. This is because although messages that initiate a sub-discussion are important, this is not the case for all initial messages. Columns 9-10 follow the same trend for the criterion of popularity.

On the other hand, the results presented in Table 4.6, in columns 6-7 and 9-10, indicate that when we choose the messages that have had reactions which contain opinions or those whose reactions include a variety of opinion polarities, we achieve higher precision than recall. This is due to the fact that the forum messages that satisfy these criteria are usually very few.

From Table 4.5 and Table 4.6, we notice that the *precision* is on average low independently of the criterion used. The low precision shows that whatever criteria we use we will retrieve non-interesting messages among the interesting ones. This is inevitable because a message that is considered key for one user may be skipped by another one and vice versa. Also, when an expert classifies the forum messages, the fact of reading the message A before the message B, affects his/her decision on which message is a key message. Even if, under certain circumstances, they would be both regarded as key, the expert tends to select the one that appeared first.

This concept is mentioned in [Rob77] regarding the Probability Ranking Principle. According to this, the fact that a document A is retrieved before a document B may affect the usefulness of A. The same is also pointed out in [ZCT02], a paper regarding redundancy detection, according to which assigning a document as redundant depends on what documents the user has already seen. Since redundancy is not symmetric, whether a document is redundant or not depends on the order of presenting the documents.

Whether we are satisfied with a high recall or a high precision is mainly a parameter of what we want to achieve. If our objective is to retrieve only interesting messages, then a high precision is desirable. On the other hand, if we want to retrieve all of the interesting messages, then we have to focus on having a high recall.

In our case, we are looking for a solution that results in a balance between recall and precision. From Table 4.5 and Table 4.6 we notice that the F1-measure gives similar average results for all criteria, with the criterion “order” having the minimum value and the criterion “reply” having the maximum. If we look at each result individually we can identify the differences in results caused by the different user attitudes. For example, in forum 2 the “reply” criterion gives a much lower F1-measure for the first user compared to the second one. This observation leads us to the conclusion that user profiling and personalization could improve the recommendation of key messages to each user individually.

The F1-measure values lead us to the following ranking of the criteria in descending order of value: reply, popularity, opinion, entropy, root and order. As a result, we can assume that the decision of a user who classifies forum messages as key or not is based on “local” information such as whether a message initiates a sub-discussion or whether it contains opinion, as well as on “link” information such as whether it has had various opinion reactions.

The low F1-measure results of each criterion separately show the need of aggregating the different criteria in order to achieve a more satisfactory coverage of the users’ needs.

4.5.5 Aggregation of Criteria

Each criterion gives different results. As we saw previously, the best criterion seems to be the “reply”, followed by the “popularity”, the “opinion”, the “entropy”, the “root” and the “order” criterion. The difference between the highest average F1-measure results and the lowest average F1 results is quite low (0.08), thus, we will not differentiate between the criteria. We will assign to all of them the same weight (=1) when it comes to their aggregation.

Initially we will normalize all the criteria so as to have values between 0 and 1. There are many ways in which multiple criteria can be aggregated. One way is to take a linear combination of all criteria by applying the same or different weights per criterion. Other ways are mentioned in [AT05] and they include Pareto optimal solutions and the successive concessions approach. These techniques differ in the way they prioritize the optimization of each criterion. For this dissertation, we choose to do a simple linear aggregation, considering that each weight equals to 1. The linear aggregation gives ranked results. Therefore, the precision will have more sense if we calculate the precision at a cut-off value n of the ranking [SL68]. In our case, n will be the number of the user's answers.

Hence, if a user has assigned n postings to represent key messages, we have:

$$recall = \frac{correctly\ assigned\ key\ messages}{n} \text{ and}$$
$$precision@n = \frac{correctly\ assigned\ key\ msgs\ in\ n\ ranked\ msgs}{n}.$$

We give the results of *precision@n* and *recall* per forum and per user in Table 4.7. We consider n to be the number of messages the user has said to be the interesting ones and we consider a message to be "correctly assigned" by our aggregated measure if it is ranked lower than the 50% of the total messages of the forum. If, for example, a forum has 100 messages, then we consider as "correctly assigned" only the first 50 ranked.

Table 4.7: Recall, Precision@n and F1-measure results for the linear aggregation of all criteria.

Forum	Expert	Recall	Precision@n	F1
1	1	0.88	0.53	0.66
	2	0.86	0.43	0.57
2	1	0.72	0.28	0.40
	2	0.88	0.25	0.39
3	1	1.00	0.33	0.50
	2	0.73	0.4	0.52
4	1	0.8	0.4	0.53
	2	0.86	0.29	0.43
5	1	0.64	0.07	0.13
	2	0.68	0.25	0.37
6	1	1.00	0.67	0.80
	2	0.76	0.59	0.66
7	1	0.86	0.45	0.59
	2	0.7	0.35	0.47
8	1	1.00	0	0
	2	0.78	0.56	0.65
Average		**0.82**	**0.37**	**0.48**

From the Table 4.7 we see that the average F1-measure increases by 60% when compared to the F1-measure of the criteria applied separately. The high increase shows that by having a simple linear aggregation of the proposed criteria we have much better results than having one criterion every time.

Let us see how successful we would be if we were recommending messages to the user using this aggregation of criteria. We do not look at the recall because it is stable no matter how many messages we recommend to the user each time. We focus, instead, on the *precision*@n measure, where n is the number of recommended messages. Our purpose is to see the effect of this number to the satisfaction of each user. In other words, we want to see whether recommending only few messages can cover the needs of each user. The results are shown in Table 4.8 and Table 4.9 for the forums we have experimented with.

Table 4.8: Recommendation satisfaction per forum (1-4). The recommended messages vary from 1 to 100.

Msgs	Forum 1		Forum 2		Forum 3		Forum 4	
	u1	u2	u1	u2	u1	u2	u1	u2
1	1	1	1	0	0	0	1	1
2	1	1	0.50	0	0.50	0.50	1	1
3	0.67	0.67	0.67	0.33	0.33	0.67	0.67	0.67
4	0.75	0.50	0.50	0.25	0.25	0.50	0.50	0.50
5	0.80	0.40	0.40	0.20	0.40	0.40	0.40	0.40
10	0.50	0.30	0.40	0.20	0.50	0.50	0.30	0.30
15	0.60	0.27	0.33	0.20	0.33	0.40	0.27	0.33
20	0.45	0.25	0.30	0.20	0.25	0.35	0.20	0.12
25	0.40	0.20	0.24	0.16	0.20	0.28	0.20	0.24
30	0.40	0.17	0.23	0.13	0.17	0.23	0.17	0.23
40	0.35	0.15	0.18	0.10	0.13	0.18	0.13	0.18
50	0.32	0.14	0.20	0.12	0.10	0.14	0.10	0.14
60	0.27	0.12	0.22	0.12	0.08	0.13	0.08	0.12
70	0.24	0.10	0.21	0.10	0.09	0.11	0.07	0.10
80	0.21	0.09	0.21	0.09	0.08	0.11	0.06	0.09
90	0.19	0.08	0.20	0.09	0.07	0.10	0.06	0.08
100	0.17	0.07	0.18	0.08	0.06	0.09	0.05	0.07

In Tables 4.8 and 4.9 the column "Msgs" shows the number of messages that are recommended to the user. From these tables we can see that when the number of recommended messages increases, the satisfaction of the user

Table 4.9: Recommendation satisfaction per forum (5-8). The recommended messages vary from 1 to 100.

Msgs	Forum 5		Forum 6		Forum 7		Forum 8	
	u1	u2	u1	u2	u1	u2	u1	u2
1	1	1	1	1	1	0	0	1
2	0.50	0.50	0.50	0.50	1	0	0.50	1
3	0.33	0.67	0.67	0.33	1	0.33	0.33	0.67
4	0.25	0.75	0.50	0.50	0.75	0.25	0.25	0.75
5	0.20	0.60	0.40	0.60	0.80	0.20	0.20	0.60
10	0.10	0.30	0.20	0.60	0.70	0.40	0.10	0.50
15	0.07	0.33	0.20	0.67	0.53	0.33	0.07	0.40
20	0.05	0.30	0.15	0.55	0.45	0.30	0.05	0.40
25	0.04	0.24	0.12	0.48	0.44	0.32	0.04	0.32
30	0.03	0.27	0.10	0.43	0.40	0.37	0.03	0.27
40	0.05	0.23	0.08	0.40	0.33	0.35	0.03	0.23
50	0.06	0.20	0.06	0.32	0.28	0.30	0.02	0.18
60	0.07	0.22	0.05	0.28	0.25	0.27	0.02	0.15
70	0.06	0.20	0.04	0.24	0.23	0.27	0.01	0.13
80	0.05	0.21	0.04	0.21	0.20	0.25	0.01	0.11
90	0.04	0.20	0.03	0.19	0.19	0.23	0.01	0.10
100	0.05	0.18	0.03	0.17	0.18	0.22	0.01	0.09

decreases in most cases. This is because only a subset of the messages of a forum can be interesting for a user, and usually this subset contains few messages.

In Table 4.10 we present the average satisfaction values per set of recommended messages. We can see that the more the recommended messages, the less the precision. Hence, in order to satisfy the user we have to recommend few messages. The results are independent of the total number of forum messages.

Table 4.10: Average recommendation satisfaction measured by precision@n per set of messages ranging from 1 to 100.

Messages	Average
1	0.69
2	0.63
3	0.56
4	0.48
5	0.44
10	0.37
15	0.33
20	0.27
25	0.25
30	0.23
40	0.19
50	0.17
60	0.15
70	0.14
80	0.13
90	0.12
100	0.11

We observe that by using the aggregation of the proposed criteria we can have a satisfactory recommendation of messages to a user. It has to be noted that we have used a linear combination of the criteria where each weight equals to 1. In the future, experimenting with different weights can optimize the aggregation results.

4.5.6 Additional Experiments

In this Section, we present the second part of the experiments we have carried out in order to see whether the messages that we extract are useful when they are recommended to the users. Initially we select some forums and we recommend few messages of each forum to some users. The recommended messages are the ones we extract when we aggregate the presented criteria. Then we see whether the users are satisfied with this recommendation.

We performed the experiment with 6 users, 8 French forums and 7 English forums. We evaluated a total of 35 answers (almost 6 forums per user). For each evaluation we started by presenting the forum to the user by giving its title and a list of the 20 first extracted messages given in random order. We asked the users to read each of the recommended messages separately and rate them according to how useful they are in helping the user to get an idea of the forum's content and start navigating inside it. The rating varied from "useful", "low usefulness/indifferent" and "useless". The explanations for each rating are presented in Table 4.11 and it is exactly how they were given to the users.

Table 4.11: The explanations of the ratings as given to the subjects of the experiment.

Useful	A message for which you would be interested to see what it replies to or which replies it has received.
Low usefulness/indifferent	You are not interested in this message and you do not really mind about the discussion that took place before or after it.
Useless	This message does not help at all in getting an idea about the forum's content and you are also not interested in how it has been replied.

The outcome of this experiment is shown in Figure 4.7 where we see the percentage of "useful", "indifferent" and "useless" to the users messages in

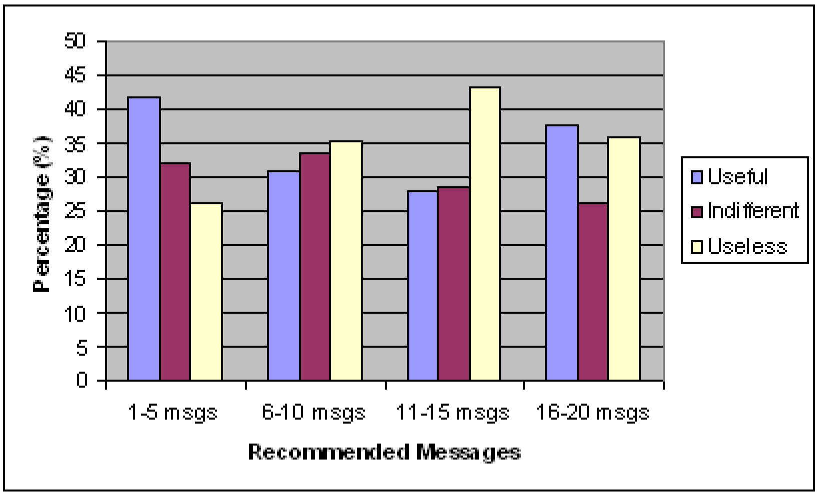


Figure 4.7: Results of the evaluation of a recommended set of messages.

the set of the first five recommended messages, the set of the second five recommended messages and so on.

This experiment showed that within the first 5 extracted messages we find the maximum of "useful" and the minimum of "useless" postings. At the same time when we recommend more than 10 messages to the user, we include a lot of useless ones.

During our experiment, we noticed that some short messages may have been popular but they made no sense to the user when they were presented on their own. As a result we decided to carry out another experiment after having removed these short messages from the set. This improved the results of the evaluation as it is shown in Figure 4.8, by increasing the rate of useful messages by an average of 6% and reducing the rate of useless messages by an average of 7%. When the short messages are excluded, we can see that the number of useless messages lowers even when we go up to 20 recommended messages, while the number of useful messages remains high.

During the experiments we noticed that there are differences in how users rate each posting. For example, a "useful" posting for one user was sometimes rated as "useless" by another one. This shows that there is a need for personalization techniques which will improve even more our results and will lead to a more appropriate recommendation per user. This agrees with the comments we have had from the users themselves during the experiment

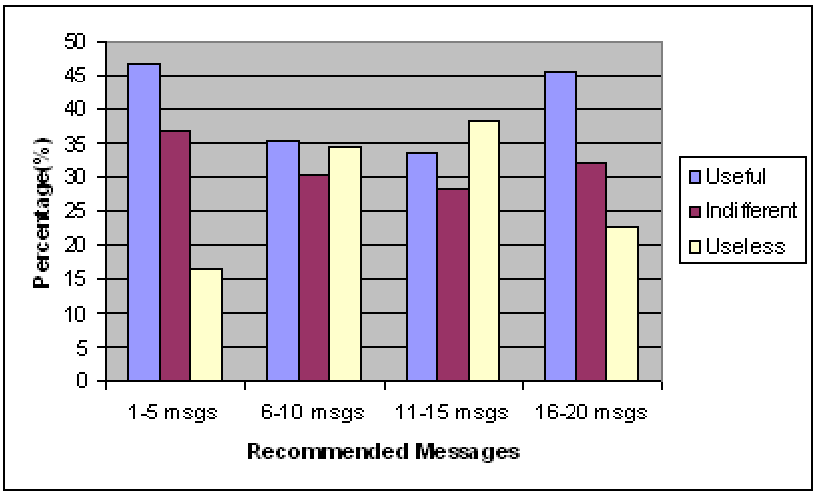


Figure 4.8: Results of the evaluation of a recommended set of messages when the short messages are excluded.

which are summarized in the following points:

- "My beliefs affect my rating": if a posting is in favor of an idea, political or not, for which the user is totally against, then the posting is more likely to be rated as "useless" because the user is bothered and s/he does not want to know more about this posting. The opposite stands as well, in the sense that a posting that complies with the beliefs of a user will more probably be considered as "useful".

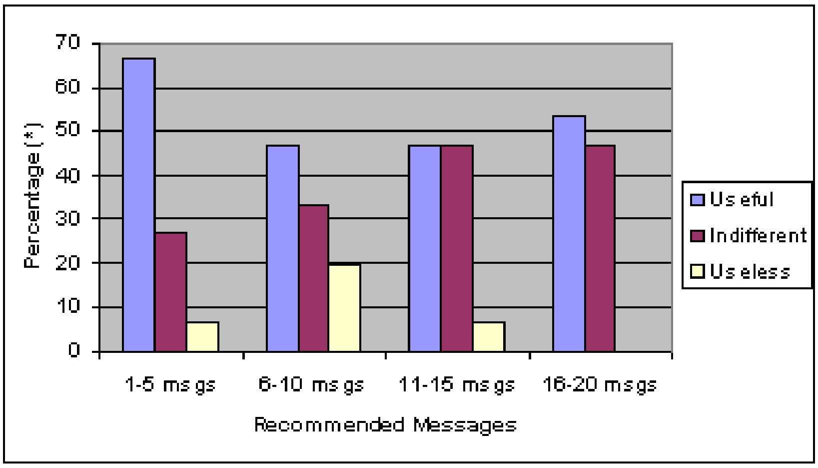
- "I am already informed about this issue and I do not want to lose time reading more about it": a posting that contains information which the user already knows about may be rated as "useless" or "indifferent".

- "I am not familiar with the language used in this posting": a posting that uses high level vocabulary is often not considered as useful by the users. In addition, if the user's mother language is not that of the language of the posting, the user may not completely understand the content and as a result s/he may rate a posting as "useless" even if it is not.

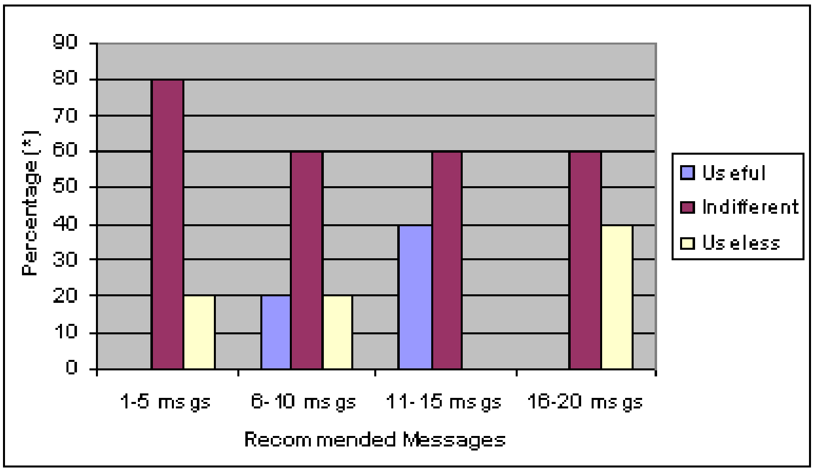
- "I prefer the aggressive postings": for some users the style of a posting plays a role in what they consider as "useful". For example, postings that contain slang are attractive to some users but not to others.

The results presented in the tables and figures of this Section are the total results of all the 15 forums together. We have to note, though, that the results can differ significantly according to the forum. For example in Figure 4.9 we see that for the forum presented in (a) most proposed postings were rated as useful, while for the forum presented in (b) most proposed postings were rated as useless. During a more elaborate research inside each of these forums, we saw that indeed the first forum was consisted in interesting postings, while the second one was generally not an interesting forum and, thus, no matter which postings we recommended they would be indifferent to the user. This is a general problem in the evaluation of recommender systems in the sense that even if the system does its best, if none of the available postings are interesting to the user anyway, the performance will be low.

The reason why the second forum shows so bad results is that it consists of a lot of small messages that do not make sense on their own but only as part of the thread they are in. This is a very particular case of a forum. In the future, we should investigate how to treat such forums with small messages. Probably the recommendation of the whole discussion thread or chain instead of messages on their own could be a solution.



(a)



(b)

Figure 4.9: Results of the evaluation of a recommended set of messages for two different forums; one forum contains a lot of "useful" messages while the other one contains very few "useful" postings.

4.6 Conclusion

In this Chapter, we have used the information extracted from the Post-Reply Opinion Graph in order to retrieve and classify the discussion postings according to how key they are for the discussion. These key messages can be recommended to users in order to help them browse through an online discussion without having to read all the existing postings.

We have experimented with real users and forums in both English and French. Various criteria have been applied and an aggregation of these criteria seems to improve the results of the recommendation task. Recommending messages to a user is quite a subjective issue and as a result personalization techniques should be applied in the future in order to make more appropriate per user recommendations.

Chapter 5

The System Prototype

In this Chapter we describe features and functionalities of a system prototype that has been implemented as part of this thesis. The prototype implements the proposed model together with structure, opinion, topic, time and user-oriented measures. The recommendation approach mentioned in Chapter 4 is also included. The prototype allows us to evaluate the analysis of forums found on the Web after having been parsed and inserted into a local database.

5.1 Introduction

For the purpose of evaluating our model, we have developed a system prototype which demonstrates proof of concept. The system aims at showing how our proposed framework facilitates the navigation of a user within an online discussion and the extraction of useful information from it.

The prototype gives the user the opportunity to quickly access interesting information within small or large online discussions. We use online discussions in the form of web forums such as the forums from the site of http://www.liberation.fr/, an example of which appears in Figure 5.1. In such forums, the users use a pseudo name in order to identify themselves and they can send new messages or reply to existing ones. The indentation in the forum page points out “reply-to” links.

Figure 5.1: A web forum as it appears on the site of http://www.liberation.fr. The indentation implies "reply-to" links. The user names have been replaced by grey squares for privacy reasons.

The prototype enables the browsing inside a forum through a visualization tool which offers various functionalities and access to both the social network and the Post-Reply Opinion Graph of a forum. In Figure 5.2 we can see a screenshot of the implemented tool which is called "Discussion Analysis Tool".

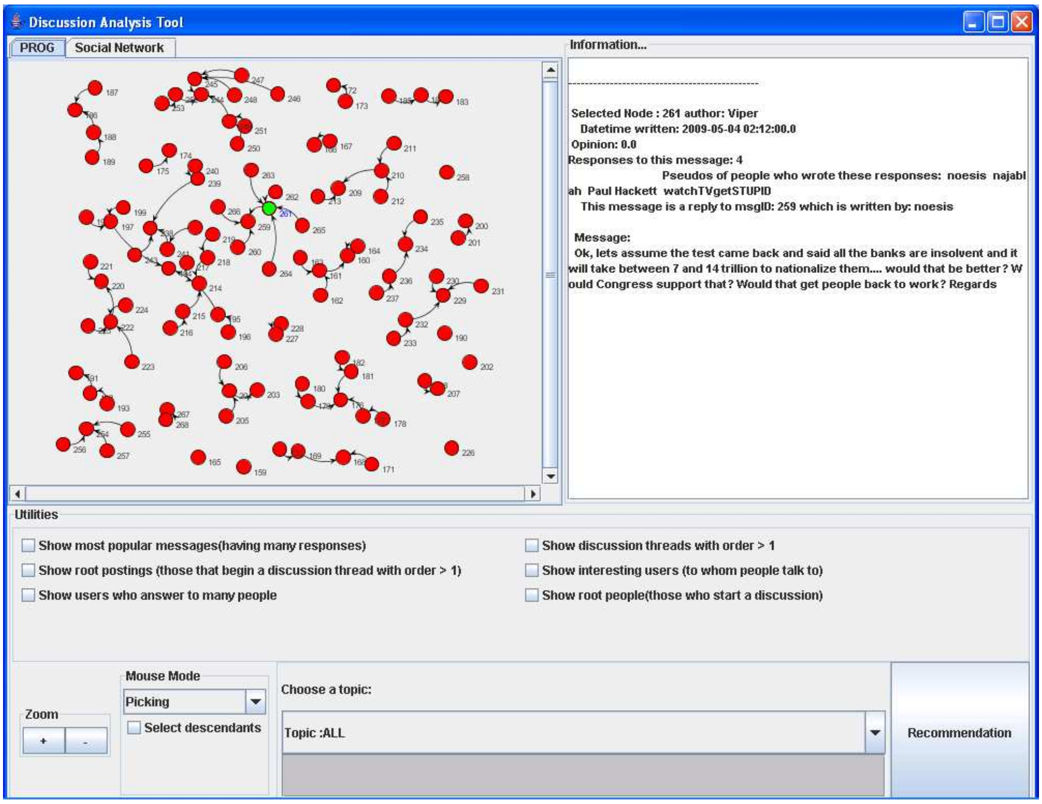


Figure 5.2: A screenshot of the system prototype, the Discussion Analysis Tool.

The main frame window of the Discussion Analysis Tool is divided into three parts. One part is reserved to the graph representation of the social network of a forum and the Post-Reply Opinion Graph model. Another part consists of a utilities panel which allows the user to mine the forum by selecting vertices which satisfy certain criteria, navigate through the various sub-topics of the forum and get a recommendation of messages to start-with. The third part of the main window is a text area which permits information to be printed regarding selected vertices and selected topics.

Part of the prototype functionality has been developed for the project

"Conversession" which is a project for a start-up company supported by CREALYS. The objective of this project is the automatic analysis of forums and the direction of a moderator towards a subset of the forum's messages. The project aims at helping the moderator to take decisions and intervene in the discussion.

The rest of this Chapter continues with a presentation of the functionalities of the system. In Section 5.3 we emphasize on implementation issues and in Section 5.4 we present the analysis of a web forum through the Discussion Analysis Tool. Section 5.5 concludes the Chapter.

5.2 System Functionalities

In Figure 5.2 we can see the main screen of the prototype system which is divided into various areas according to the provided functionalities. On the top-left hand side, we have the visualization of the two graphs that are on a tabbed panel. By clicking on "PROG" we have the Post-Reply Opinion Graph and by clicking on "Social Network" we can see the user-based graph of the selected forum. On the top-right hand side, there is a panel where we can see information about the selected vertices of any of the two graphs. At the bottom part of the screen, we have the functionalities that we can apply to either graphs, such as the possibility to zoom in and out the graph, to move it inside its reserved window, to select a topic and see the postings which belong to it and to get recommended messages.

A summarized list of the system functionalities is presented below:

Social Network (user-based) graph visualization and navigation. The parsing of a forum results in a user-based graph which describes the social network of the forum from the point of view of user interactions which take place. The user can select a user node which represents a user u and retrieve the information presented in Table 5.1.

Other functionalities that are applied to the social network graph include:

- the identification of the most popular user,
- the identification of users that talk to many people and
- the identification of users who initiate a sub-discussion.

Table 5.1: Information that can be extracted from the system regarding a user u.

the number of people u has talked to
the names of people to who u has talked to
the number of people that have talked to u
the names of people who have talked to u
the average opinion of u inside the forum
the average opinion towards u inside the forum
the messages u has written in a specific topic if this topic is selected
the average opinion of u inside a topic if this topic is selected

By clicking on the respective checkboxes, we have the relative user vertices highlighted.

Post-Reply Opinion Graph visualization and navigation. The parsing of the forum results in a second graph: the Post-Reply Opinion Graph. The user can select a vertex which represents a posting and extract information such as that presented in Table 5.2.

Table 5.2: Information that can be extracted from the system regarding a posting.

the author who has written the posting
the date and time that shows when the message was posted
the opinion polarity included in this message
the opinion entropy of replies towards the specific posting
the average opinion of replies towards the specific posting
how many replies did this message have
the names of the users that have replied to this message
to which message this posting replies to, if applicable
the content of the message

Other functionalities that are applied to the Post-Reply Opinion Graph include:

- the identification of the most popular messages,
- the identification of the discussion threads with an order greater to 1,

- the identification of postings which initiate a discussion thread,
- the first and the last posting of the forum,
- the postings with a high opinion entropy in their replies,
- the identification of the descendants and ancestors of a vertex as this is given by the structure of the graph and
- the identification of the ancestors of a vertex regarding the time and the topic without necessarily respecting the graph structure.

Likewise the social network graph, the vertices that satisfy the selected criteria are highlighted on the graph.

Interaction between the Social Network and the Post-Reply Opinion Graph. Both graphs are concurrently represented on the screen of the Discussion Analysis Tool. By selecting a user vertex from the social network graph, the postings written by this user are highlighted in the Post-Reply Opinion Graph. In this way, we can see how often the user has appeared inside the discussion and in which discussion threads. We can also see whether the discussion threads where this user has participated contain many other postings, how many times the user has posted messages in the particular thread, and whether other people have been interested in these postings.

Combination of topic identification and graph visualization. The application of a topic classification tool allows the identification of the different topics present in the forum and the assignment of each message to the identified topics. By selecting a topic, we can see the discussion threads from the Post-Reply Opinion Graph which contain postings that belong to the selected topic. Also, in the social network graph, the vertices of the users who have written messages that belong to the selected topic are highlighted.

Message Recommendation. The user can select to see some of the messages that are considered to be interesting inside the forum. The messages are selected by the method given in Chapter 4 and their respective vertices are highlighted on the Post-Reply Opinion Graph. In this way, the user can start navigating a new forum by these messages and their discussion threads. The number of recommended messages presented to the user is defined in the parameters of the system and it can vary according to the user requirements.

5.3 System Implementation

The prototype is entirely written in Java (JDK Compliance 5.0). For its implementation, we have used the JUNG java library (http://jung.sourceforge.net), the MySQL database, the SentiWordNet 1.0 Opinion Mining resource [ES05b], the TreeTagger tool [Sch94] and the topic-extraction AGAPE system [VG07].

The technical implementation of the system prototype consists of the following issues:

Database: Design and development of a relational database that stores the information extracted from web forums. The relational database contains information regarding the discussion, the users and the postings. The database schema is depicted in Figure 5.3. The table "DISCUSSION" holds general information about each online discussion, while the table "MESSAGE" holds the postings, the table "USER" the users that participate into a discussion and the table "TOPIC" information about the identified topics that occur inside each discussion.

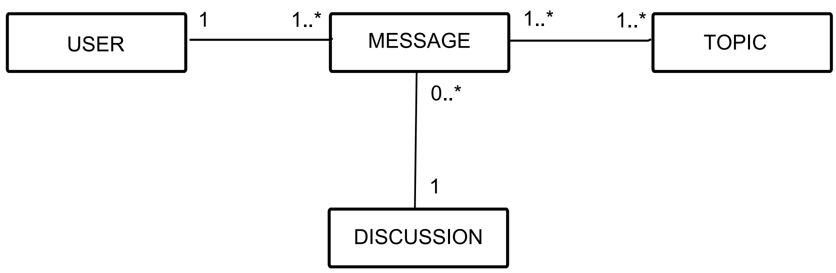


Figure 5.3: The database schema of the prototype system.

Opinion Mining: Application of Opinion Mining techniques in order to identify the opinion polarity held by a message. The prototype can represent forums of any language whose "reply-to" relations and opinion orientations are known. For the forums written in English, the opinion polarity is automatically identified with the use of the SentiWordNet resource [ES05b] described in Chapter 2. We have entered the information included in this resource into a database for better manipulation. The procedure followed in order to identify the opinion polarity per English message is the following:

1. Tokenization and part-of-speech tagging of each message by applying the TreeTagger tool [Sch94].
2. For each noun, adjective, verb and adverb, get the positive and the negative value from the SentiWordNet database. Since no Word Sense Disambiguation takes place, we consider all senses of each word. We sum up the positive values and we extract the negative ones. If the result is greater than 0 then we assign the polarity of the word to be positive (1), if it is less than 0, we consider the word to be negative (-1), otherwise we consider it objective (0).
3. If the majority of words in a message are positive, then we consider the message to have a positive polarity. For all other cases, the message is assigned to have a negative polarity unless no subjective word is contained in it.

For example, the polarity of the message "It is war." will be assigned a negative polarity. The only subjective word is the noun "war". The senses captured by the SentiWordNet tool for the instance of the word as a noun are 4 and they have a total of 0 positive values and a total of 0.25 negative values. Since the negative is higher than the positive, the word is assigned a negative polarity and as a result the message becomes negative as well.

For the forums in other than the English language, the opinion identification is done manually for the time being.

Post-Reply Opinion Graph: Construction of a Post-Reply Opinion Graph for the representation of each forum. The Post-Reply Opinion Graph has been implemented as a *DirectedSparseGraph* whose vertices are of type *DirectedSparseVertex* and each edge is a *DirectedSparseEdge* object. Each vertex represents a posting and it holds the following attributes:

- a unique message id,
- the message itself,
- the author of the message,
- the date and time which shows when the message has been written and

- the opinion polarity contained in the message implemented as a float variable.

The procedure that shows the construction of the PROG graph is shown in Figure 5.4.

```
procedure createPROG (id_debate)
    DirectedSparseGraph g;
    MessageVertex mv, mvRep;
    for all mv in DB (id_debate) do
        mvRep = getReply (mv);
        if (mvRep == null)
            mv = new MessageVertex(as root);
        else
            mv = new MessageVertex(mvRep);
            edge = g.addEdge(mv, mvRep);
    done
    end procedure
```

Figure 5.4: The construction of a Post-Reply Opinion Graph g. The messages are extracted from a database DB.

Extraction of messages by criteria: Utilization of the Post-Reply Opinion Graph with the purpose of identifying different types of messages e.g. "root" messages, "popular" ones, messages that contain positive or negative opinions etc.

Recommendation of messages: Utilization of the Post-Reply Opinion Graph in order to recommend messages to users. The recommendation approach is the one described in Chapter 4.

Social Network: Construction of the social network of a web forum in the form of a user-based graph. This graph can be extracted by the Post-Reply Opinion Graph since the information about the user is encapsulated in the vertices of this graph.

Extraction of users by criteria: Identification of various roles of users from the social network graph such as popular users, users that send a lot of messages etc.

5.4 Analysis of a real web discussion

In this Section we will analyze a real forum extracted from the Web. The analysis is similar to the one presented in Chapter 3, but the objective of the analysis in this Chapter is to show how we can visualize it through the system prototype. The analysis of web discussions through our prototype evaluates the model itself.

For the demonstration of the prototype system, we select a forum from HuffingtonPost.com (http://www.huffingtonpost.com), an English-speaking news site. The title of the forum is "NATO jets bomb fuel tankers; Afghans say 70 killed". The forum is in English and it consists of 228 messages. The users who have participated are 118.

For this forum, the opinion polarities have been identified by SentiWordNet 1.0 [ES05b]. The Post-Reply Opinion Graph appears in our prototype system as in Figure 5.5. The message objects appear with an identification number calculated internally by our application.

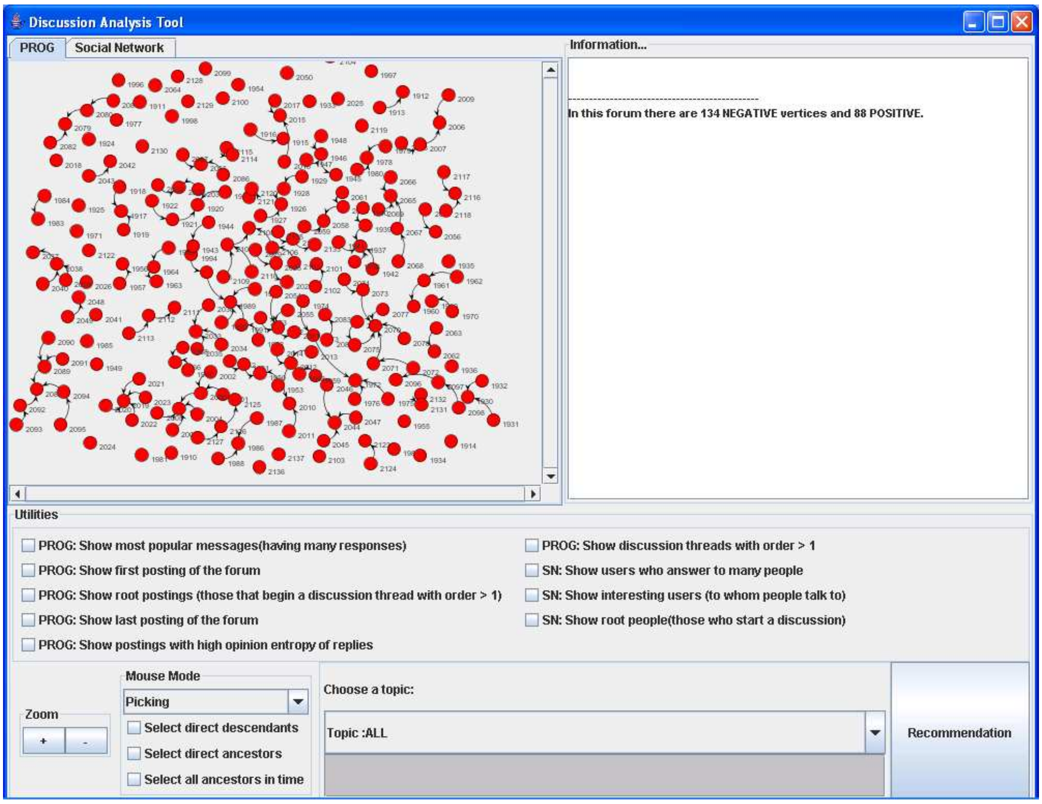


Figure 5.5: The Post-Reply Opinion Graph of the English Web forum, as it appears on the system prototype. The number of negative and positive postings appears on the right-hand side information panel.

Let us see how the prototype system can facilitate the forum analysis by describing some actions followed by a user named Bob who has launched the system with the objective to navigate within the particular forum.

General opinion atmosphere of the forum. Bob launches the system and immediately he can see, on the right-hand side of the window (Figure 5.5), at the "Information" panel, information about the opinion atmosphere of the forum. For the particular forum, he is informed that there are 134 postings with a negative polarity and 88 postings that are positive-oriented. Hence, Bob realises that the majority of the forum discusses in a negative way.

Duration of the forum. Bob wonders with what posting this forum started and whether it has been active for a long time. He clicks on the option "PROG: Show first posting of the forum". As shown in Figure 5.6, the vertex which represents the very first posting is highlighted on the graph. By reading the content of the first posting, he sees that the discussion has started with a reference to troops. Then, he clicks on "PROG: Show last posting of the forum". As shown in Figure 5.7, the vertex of the last posting is highlighted. By the information he gets on the "Information" panel, he can calculate the duration of the forum. In the particular forum, the duration has been about four days.

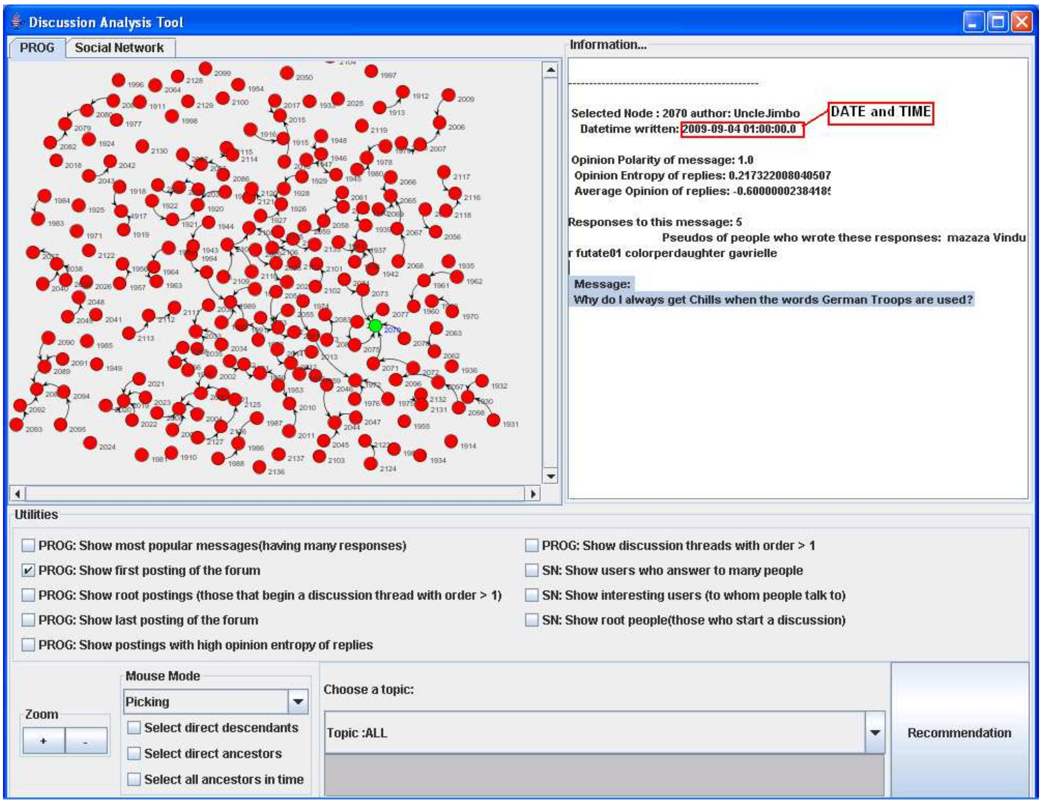


Figure 5.6: The very first posting of the forum is highlighted when the respective option is selected. On the right-hand side panel, information about this posting appears.

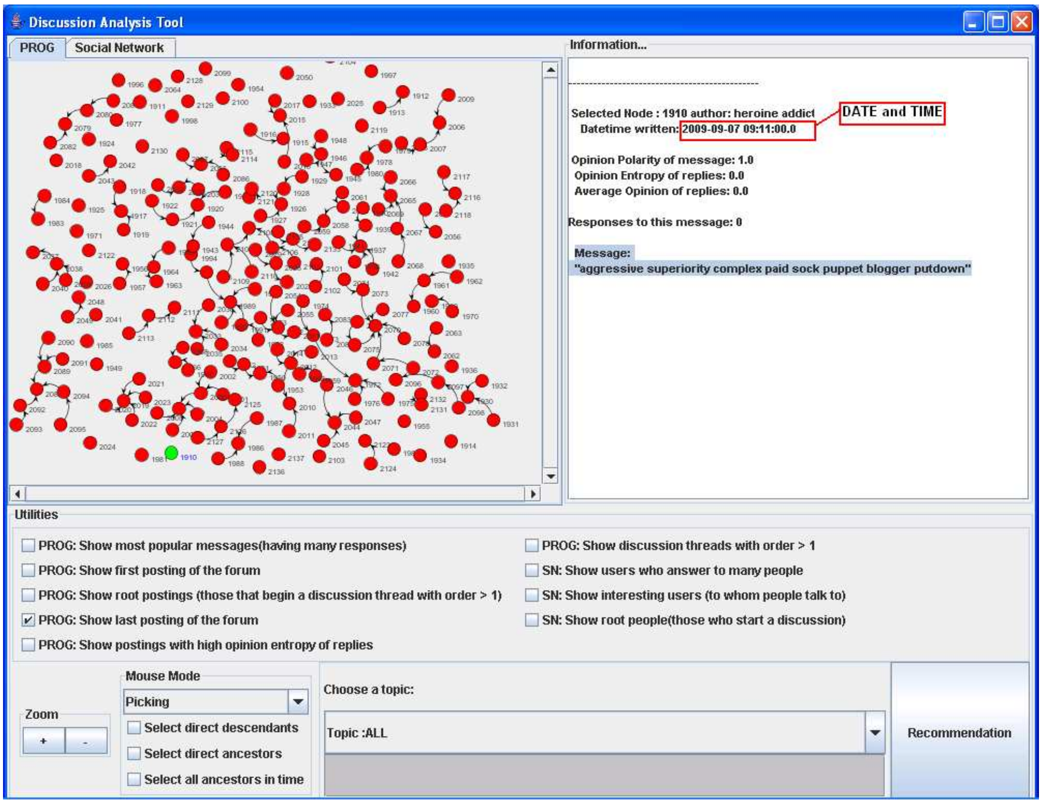


Figure 5.7: The last posting of the forum is highlighted when the respective option is selected. On the right-hand side panel, information about this posting appears.

Initiation of discussion threads. Bob wants to see which messages have initiated discussion threads so that he can choose which of the discussion threads may interest him and have a look into them in more detail. He clicks on the checkbox "PROG: Show root postings", and as we see in Figure 5.8, he can see the vertices that have started a discussion thread with more than one postings. By clicking on these root messages, he can see their message content at the "Information" panel on the right hand side. For the specific forum, he sees that discussions go around Obama, NATO, the situation in Afghanistan in general, etc.

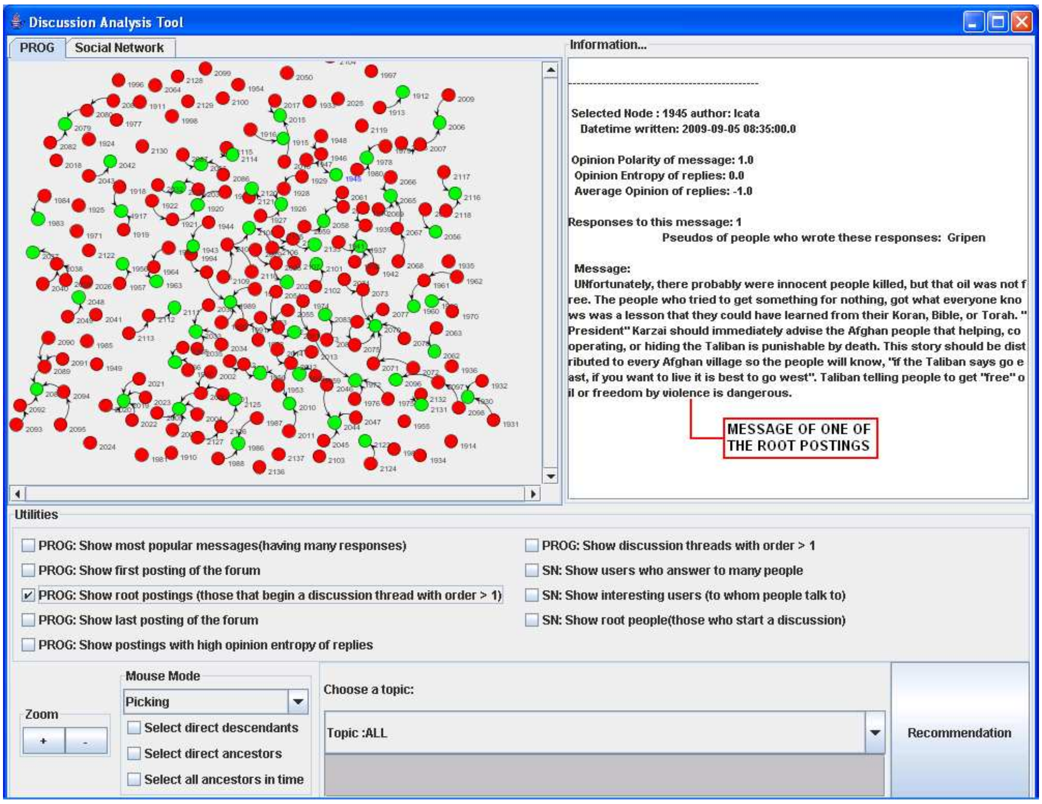


Figure 5.8: The root option is selected and the vertices that represent "roots" are highlighted.

Message Recommendation. By clicking on the button "Recommendation", Bob gets a predefined number of PROG vertices highlighted. These are messages that the system has chosen as interesting messages for a user to start-with. From the recommended messages he sees that people compare the reactions towards the Afghanistan situation and the American one.

Popular messages. Bob decides to start navigating the forum by the most popular message. In order to identify the most popular message(s), he clicks on the "PROG: Show most popular messages". The particular discussion has one message that is more popular than all the others. This is the vertex that appears with a different color (green in the prototype) than the others in Figure 5.9. Bob looks both at the graph and the "Information" panel, and he notices that the most popular message has received five replies.

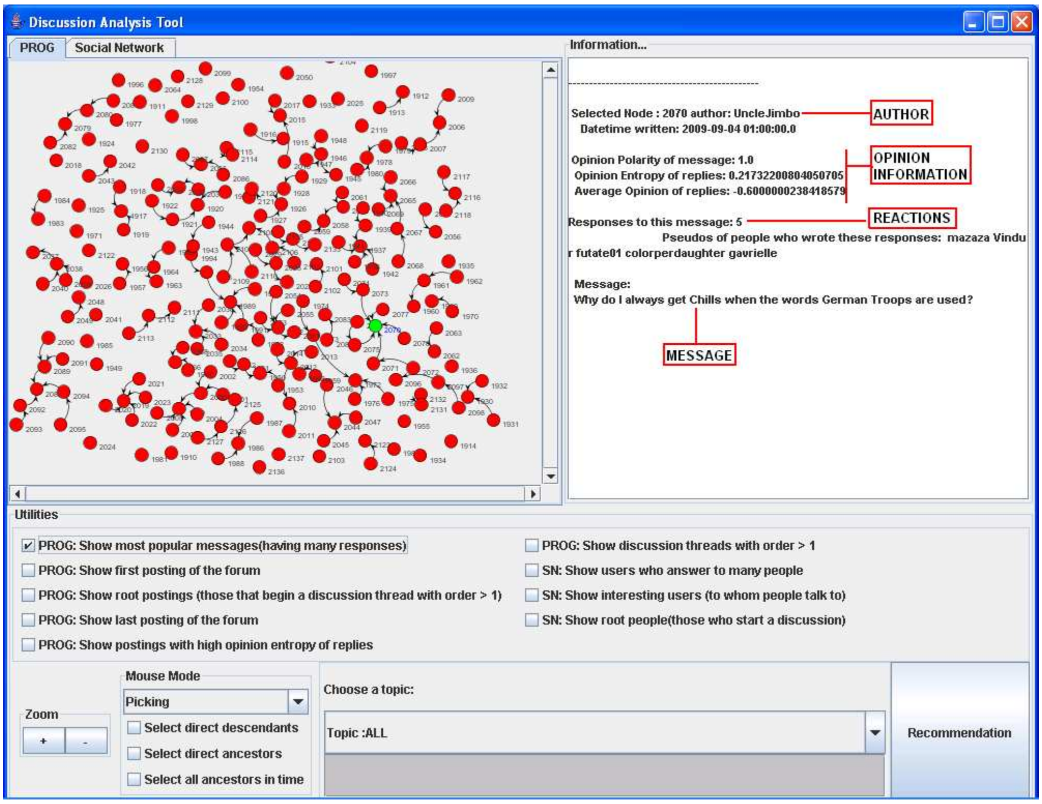


Figure 5.9: Selection of the most popular message which appears in light color (green). At the right hand side, we can see information about this message.

On the "Information" panel, Bob can see the author of this message, the opinion polarity it contains as it is identified by SentiWordNet, the average opinion and entropy of the replies towards it, whom has replied to this message and of course its content. Since he is interested in this message, Bob wants to follow the discussion around the subject evoked by this message, so he clicks on the vertices around it and he gets the information for each vertex on the "Information" panel.

Dispute inside the forum. Bob would like to know which postings have caused more dispute than others. In order to see this, he clicks on the option "PROG: Show postings with high opinion entropy of replies". By reading the content of the highlighted postings, he can see that the postings that have caused more dispute are those that talk about Obama's choices, references to other countries situation, and opinions about killing.

Topic Selection. By clicking on the bar that specifies the topics, Bob can see the topics that appear in the particular forum. In this forum, the main topics include discussion about war, deaths, enemies, family. Every time a topic is selected, the PROG graph changes and it shows the messages that belong to the topic together with their discussion threads.

An example is shown in Figure 5.10 where the topic with identifier 36 has been chosen. This is a topic that regards messages which discuss mainly about war since the main keywords that characterize this topic are: war, soldier, crime. The vertices with the lighter color (green in the prototype) represent those messages that belong to the topic. In the Social Network the users who have written messages that belong to the selected topic are highlighted. This is shown in Figure 5.11 where the user vertices that are lighter color (green) represent authors who have written messages in the particular topic.

On the "Information" panel, Bob can see the average opinion in this topic and moreover he can see that the negative postings are more than the positive ones.

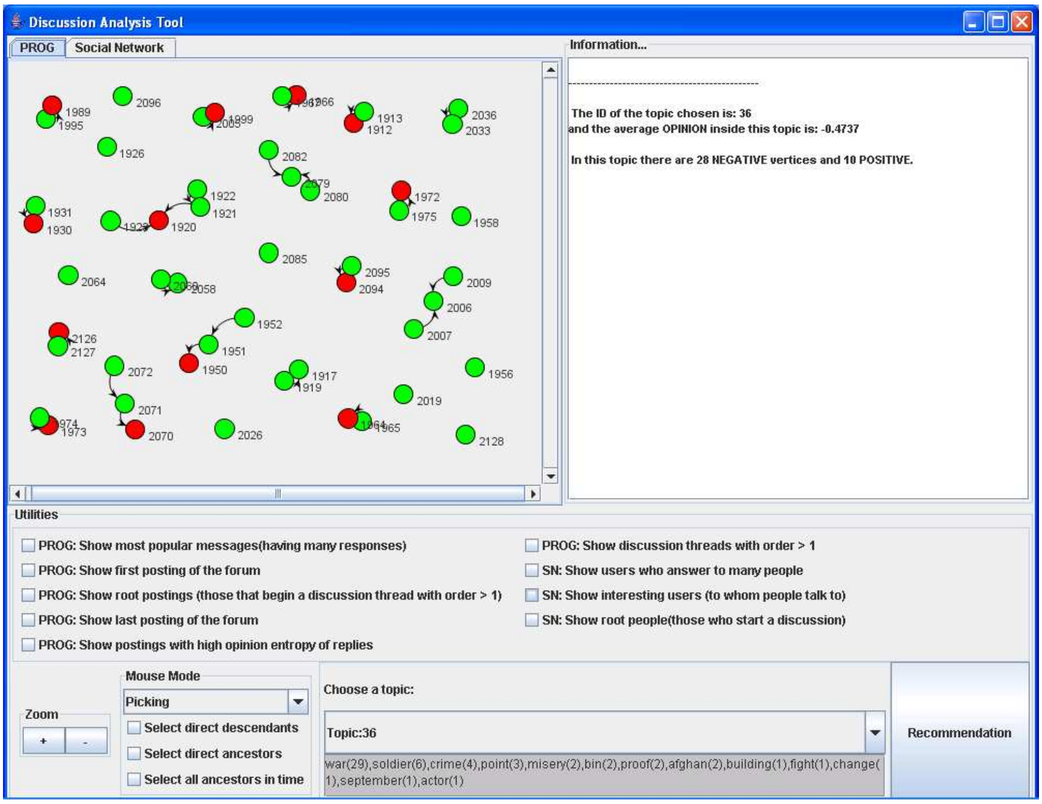


Figure 5.10: A subset of the PROG graph specific to the topic selected. The light color (green) vertices represent postings that belong to the selected topic.

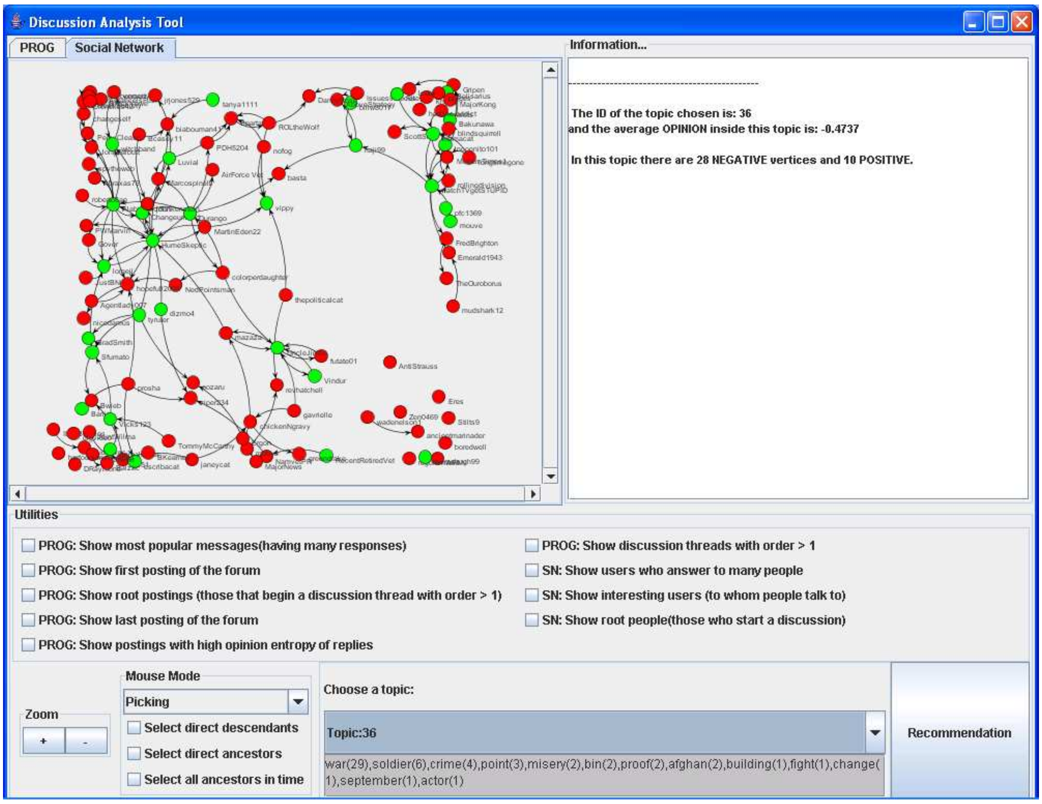


Figure 5.11: The social network of the users of the forum. In light color (green), the user vertices which represent users who have posted at least one message that belongs to the selected topic.

Ancestors of a posting. Bob wants to see all the postings that have preceded a specific message. In the "Mouse Mode" panel, which is located at the bottom side of the prototype's screen, he can choose what else to select when a certain vertex is selected. Regarding the ancestors, there is the option between "Select direct ancestors" and "Select all ancestors in time". The first option shows the chain of the ancestors of a posting as it is expressed through the edges of the graph. The second option shows the ancestors of a posting irrespective of the graph structure but based on the temporal dimension. The option is shown in Figure 5.12. By looking at the messages that have preceded a specific message, Bob can see which messages may have influenced a posting. In case a topic is selected, then the ancestors are shown related to this topic.

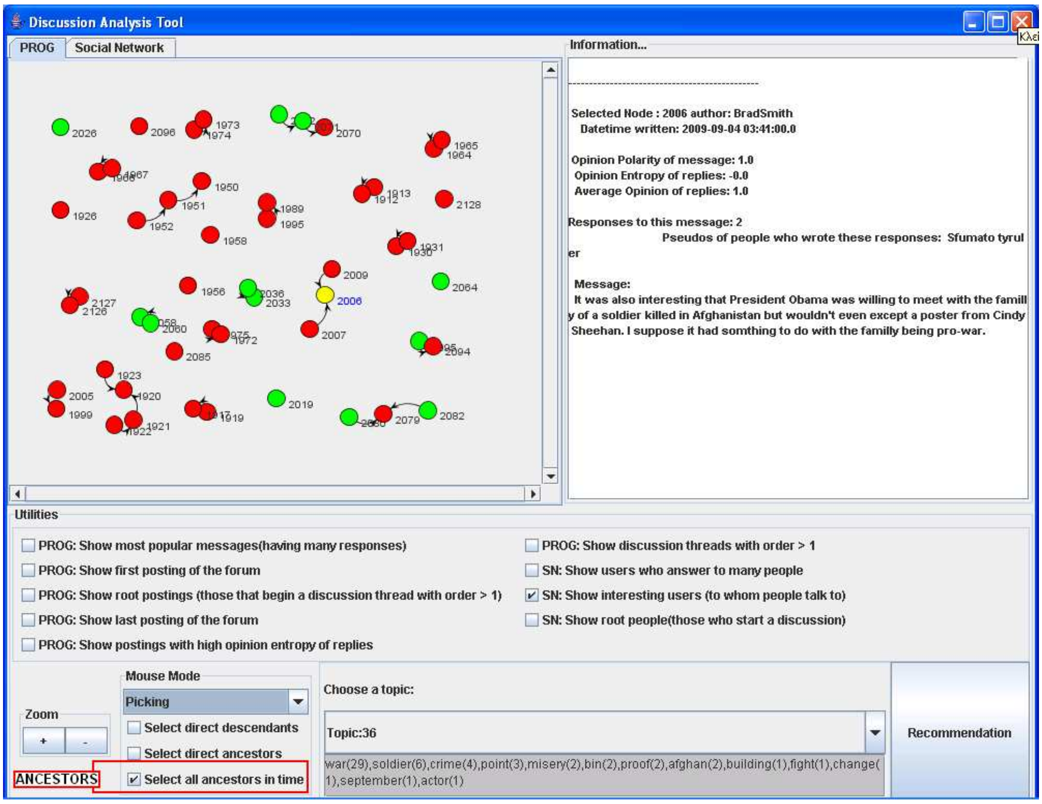


Figure 5.12: The option that allows seeing all the ancestors in time of a selected posting.

Interesting users. Bob wants to know what the most popular user of the forum is talking about. By going to the "Social Network" and clicking on the "SN: Show interesting users" option, he gets the vertex of the most popular user highlighted. By clicking on this vertex, he can see the messages this user has written, as these are highlighted on the PROG graph. By these messages, Bob knows that the most popular user discusses about al Qaeda and Taliban.

5.5 Conclusion

In this Chapter we have presented the system prototype that has been implemented in order to automate the analysis of a forum. Through the analysis of a real web forum, we have seen what kind of information we can extract.

The Discussion Analysis Tool has been implemented in such a way that makes it extendable and easily modified. It can be enriched with extra functionality and it can be adapted according to specific needs of users. According to the user's role, whether this is a moderator, a discussion analyst or simple user, the tool can be modified in order to focus more on the individual user needs.

Chapter 6

Conclusion and Perspectives

6.1 Thesis Summary

The increasing availability and the dynamic evolution of online discussion systems such as forums and newsgroups, point out the need for analysis and mining of the data that reside in them. In this thesis, we have presented a novel model for the representation of online discussions. The measures proposed together with an application of interesting messages retrieval and recommendation allows the monitoring of the textual information flow and the analysis of a discussion.

MODEL and MEASURES. Our main contribution is the proposal of a graph-oriented model for the representation of online discussions. Current research works view the online discussions as a social network between users, and, thus, they represent them from a user-oriented point of view. The representation of a discussion as a social network focuses on the participants and the interaction between them. Important information such as the structure of the discussion and the content is lost.

The proposed model presented in this thesis, goes beyond the exploitation of the developed user network. It emphasizes on the structure of the discussion and on the content from a topic- and opinion-oriented point of view. It combines Text and Opinion Mining techniques with Social Network Analysis concepts. It detects, therefore, how the structure of the discussion affects the roles of a posting (initial, popular, high opinion entropy, etc.) and also the relations between the different postings (e.g. detection of ancestor postings that may have influenced the sending of a particular posting).

The definition of the Post-Reply Opinion Graph on which our model is based allows the proposal of measures that aid in the analysis and the mining of online discussions. These measures use structural, opinion, temporal and topic information in order to facilitate the analysis of a discussion.

RECOMMENDATION. An application of our model is the extraction of interesting messages from a discussion and their recommendation to a user. This can facilitate the browsing of a user inside the discussion by proposing a set of key messages to start with in order to identify the content of the discussion. A number of criteria have been studied based on the structure and the encapsulated information of the Post-Reply Opinion Graph. These criteria are correlated with user preferences. Extensive experiments have been carried out in order to see under which conditions a set of recommended messages is acceptable by a user.

PROTOTYPE. We have developed a system prototype which facilitates the representation and analysis of online discussions. It allows a user to interact with an online discussion, browse through and zoom into it. The prototype system accepts easily the integration of Text and Opinion Mining libraries and thus, it enables the representation of discussions in multiple languages.

6.2 Future Research

The introduction of the model presented in this thesis opens the way for further research and challenges. Some of them are presented in this Section.

Model Combination. One future objective of our work is to combine the two models, the social network one and the novel model, for a better discussion analysis. For example, we could extract the user-based graph from the Post-Reply Opinion Graph and identify users who satisfy a certain role [STE07]. Then, we could extract from the user-based graphs the users who are experts or influencers [DCTM08, ZAA07] in the particular discussion. Users who are authorities or experts on a given topic may have more interesting postings than users who do not have this role [HLS+07, ZAA07]. Going to the Post-Reply Opinion Graph, we could extract the respective postings and the recommendation list of interesting messages can be updated with these postings.

Similarly, we could combine the two models in order to find out whether popular users write interesting messages, how authors of popular messages influence the flow of the discussion etc.

Opinion Evolution. A posting can be considered to have an effect to all postings that follow it in the future. Using the temporal as well as the topic information we can identify the ability of a certain posting to change the existing opinion of a user.

Moreover, capturing the opinion presence could enable the identification of agreement and disagreement between messages which cannot be determined just by the orientation of each message [SC08]. In the future we are planning to carry out experiments in order to find out if and under which circumstances our model facilitates this identification.

Topic Identification. The structure of the Post-Reply Opinion Graph allows the extraction of discussion threads and chains. This knowledge could be used in the future in order to give confidence to topic identification algorithms. For example, two messages that appear in the same discussion chain or thread have higher probability to belong to the same discussion topic. As a result, a topic-identification algorithm could give higher probability of belonging to the same topic to messages that do not only have similar content but they are also linked in the same chain or thread.

Structure. In the presented work, the relations between the messages are considered to be known (which message replies to what) but often these relations are not available. We have described a way on how topic and temporal information can aid in the identification of the real ancestors/descendants of a posting. In the future, we intend to work more with the automatic extraction of these relations between postings and the population of the graph with appropriate links.

Additionally, our model captures currently cases where one message can reply to one and only one message. In some cases, though, one message may respond to more than one message. Future work needs to cater for changes in the model and the measures in order to capture this particularity.

Personalization. At the moment, our approach of recommending key messages to users is not exactly that of a recommender system. The reason is that it deals with various users for a recommending task but it does not really use the knowledge it can acquire from each user. One way to improve the recommendation task is to use personalization techniques [EV05, Mob07] and extend this work in order to have a proper recommender system.

As mentioned in [Ric79], if a user wants interesting messages, the answer depends on whom the user is. Although a user-profile with some basic user characteristics (e.g. age, gender) can easily be kept into a database, a user profile that contains information regarding what makes a message interesting for a user is not easily catered for.

Considering forums, one explicit way to find out the preferences of users regarding interesting messages would be to ask the user to rate a list of messages of a specific forum as "interesting" or not. The messages should vary between those that contain opinions, facts, question/answers, humor, or certain keywords. Having collected the ratings given by each user, we could then propose the messages rated as "interesting" to similar users (collaborative approach) or search for similar messages to propose to the specific user (content-based approach).

An implicit way that could help in developing the user profile of a forum user, would be by looking at the content of the messages this specific user has written. In this way we could recommend to the active users of the forum, messages with similar content e.g. messages of the same topic with different opinions or messages written by the same users with whom they have already discussed. Another implicit method could be to analyze the behavior of the active users. For instance, if the users click on proposed messages because they seem to them interesting, or reply to existing messages, then we can update the list of proposed messages by finding similar messages to the replied or clicked ones.

PageRank. Apart from adding personalization techniques to our approach, another interesting point would be to consider a PageRank-style criterion in order to choose the key messages. PageRank [BP98] is one of the most known link analysis algorithm which, among others, is used by the Google search engine. It has been used in order to rank web pages according to importance. It is based in the assumption that a web page that is mentioned by an *important* source is more likely to be important as well. The PageRank algorithm views the web as a graph with web pages as vertices. In order to rank by importance, it considers the *inDegree* and *outDegree* of each vertex and also the *weight* of the inDegree links i.e. if the current page is mentioned by an important one. We could apply the concept of PageRank in or ranking as well, since two messages with the same values for all criteria can be differentiated by the importance of the messages they reply to or they are replied to. This is a very interesting issue that needs further research.

Redundancy. One other way to improve our recommendation list is to cater for redundancy and exclude the redundant messages from the whole set of messages. Redundant can be a message that although it is interesting and relevant to the subject of the discussion, it contains information that has already been mentioned by messages that appear earlier in the recommendation list. The position of the message inside the message stream is an indication of whether it is redundant or not. [ZCT02] point out the same notion for a document stream. They mention that a system should identify documents that are dissimilar to the previously delivered documents in the sense of containing new information. In our case, the structure of the discussion threads and chains can help us in identifying the order of the stream. Text Mining techniques will aid in the extraction of the content and in the identification of similarities or dissimilarities between messages.

Bibliography

[ACK+05] S. Ananiadou, J. Chruszcz, J. Keane, J. Mcnaught, and P. Watry. The national centre for text mining: aims and objectives. *Ariadne*, 42, 2005.

[AG06] I. Antonellis and E. Gallopoulos. Exploring term-document matrices from matrix models in text mining. In *SIAM Text Mining Workshop, 6th SIAM SDM*, 2006.

[ARSX03] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *Proc. of the 12th International conference on World Wide Web*, 2003.

[AS06] A. Andreevskaia and S.Bergler. Mining wordnet for fuzzy sentiment: sentiment tag extraction from wordnet glosses. In *EACL*, 2006.

[AT05] G. Adomavicius and A. Tuzhilin. Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[AZ05] R.K. Ando and T. Zhang. A high-performance semi-supervised learning method for text chunking. In *43rd ACL*, pages 1–9, 2005.

[BCH05] S. Bloehdorn, P. Cimiano, and A. Hotho. Learning ontologies to improve text clustering and classification. In *29th Annual Conference of the German Classification Society*, pages 334–341, 2005.

[BH01] A. Budanitsky and G. Hirst. Semantic distance in wordnet: an experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, 2nd Meeting of the North American Chapter of the ACL*, 2001.

[BH04] S. Bloehdorn and A. Hotho. Text classification by boosting weak learners based on terms and concepts. In *4th ICDM*, pages 331–334, 2004.

[BNJ03] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Bou92] D. Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *14th COLING-92*, pages 977–981, 1992.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[BP01] C. Blake and W. Pratt. Better rules, fewer features: a semantic approach to selecting features from text. In *IEEE DM Conference*, pages 59–66, 2001.

[BS97] M. Balabanovic and Y. Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[CH04] K.B. Cohen and L. Hunter. *Natural language processing and systems biology.* Artificial Intelligence methods and tools for systems biology, Dubitzky and Pereira, Springer Verlag, 2004.

[CHS05] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.

[CLWL04] G. Cong, W. Lee, H. Wu, and B. Liu. Semi-supervised text classification using partitioned em. In *9th DASFAA*, pages 482–493, 2004.

[CMS01] M.F. Caropreso, S. Matwin, and F. Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text Databases and Document Management: Theory and Practice*, pages 78–102, 2001.

[CNZ05] G. Carenini, R.T. Ng, and E. Zwart. Extracting knowledge from evaluative text. In *3rd KCAP*, pages 11–18, 2005.

[CSW05] P.J Carrington, J. Scott, and S. Wasserman. *Models and Methods in Social Network Analysis.* New York: Cambridge University Press, 2005.

[DCTM08] M. Dascalu, E-V. Chioasca, and S. Trausan-Matu. *ASAP-An Advanced System for Assessing Chat Participants.* Artificial Intelligence: Methodology, Systems, and Applications, D.Dochev, M.Pistore, and P.Traverso (Eds.), Springer-Verlag, 2008.

[DGL94] B. Daille, E. Gaussier, and JM. Langé. Towards automatic extraction of monolingual and bilingual terminology. In *15th International Conference on Computational Linguistics*, pages 515–521, 1994.

[DL07] X. Ding and B. Liu. The utility of linguistic rules in opinion mining. In *SIGIR-07*, 2007.

[Elk07] C. Elkan. Method and system for selecting documents by measuring document quality. *US Patent 7200606*, 2007.

[ES05a] A. Esuli and F. Sebastiani(a). Determining the semantic orientation of terms through gloss classification. In *CIKM-05*, pages 617–624, 2005.

[ES05b] A. Esuli and F. Sebastiani(b). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, 2005.

[EV05] M. Eirinaki and M. Vazirgiannis. Usage-based pagerank for web personalization. In *Proceedings of 5th IEEE International Conference on Data Mining (ICDM)*, 2005.

[Fir57] J.R. Firth. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, Philological Society*, pages 1–32, 1957.

[FMR98] J. Furnkranz, T. Mitchell, and E. Riloff. A case study in using linguistic phrases for text categorization on the www. In *Working Notes of the AAAI / ICML, Workshop on Learning for Text Categorization*, pages 5–12, 1998.

[Fre98] D. Freitag. *Machine Learning for Information Extraction in Informal Domains*. Ph.D. thesis, Carnegie Mellon University, 1998.

[FSW06] D. Fisher, M. Smith, and H.T. Welser. You are who you talk to: Detecting roles in usenet newsgroups. In *Proc. of the 39th Annual HICSS*. IEEE Computer Society, 2006.

[FWRZ06] W. Fan, L. Wallace, S. Rich, and Z. Zhang. Tapping the power of text mining. *Communications of the ACM*, 49(9):76–82, 2006.

[GIS07] A. Ghose, P.G. Ipeirotis, and A. Sundararajan. Opinion mining using econometrics: A case study on reputation systems. In *ACL*, 2007.

[GNOT92] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[HDP+08] A. Harb, G. Dray, M. Plantié, P. Poncelet, M. Roche, and F. Trousset. Détection d'opinion: apprenons ls bons adjectifs! In *Atelier FOuille des Données d'OPinions (FODOP 08), en conjonction avec le 26ème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 08)*, pages 59–66, 2008.

[Hea94] M.A. Hearst. Multi-paragraph segmentation of expository text. In *32nd ACL*, pages 9–16, 1994.

[Hea99] M.A. Hearst. Untangling text data mining. In *Proc. of the 37th ACL*, pages 3–10, 1999.

[HKLJ04] J.L. Herlocker, J.A. Konstan, L.G.Terveen, and J.T.Riedl. Evaluating collaborative filtering recommender systems. *ACM TOIS*, 2004.

[HL04] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD International conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[HLL07] M. Helander, R. Lawrence, and Y. Liu. Looking for great ideas: Analyzing the innovation jam. In *KDD*, 2007.

[HLS+07] M. Hu, E-P. Lim, A. Sun, H.W. Lauw, and B-Q Vuong. Measuring article quality in wikipedia: Models and evaluation. In *CIKM '07*. ACM, 2007.

[HM97] V. Hatzivassiloglou and K.R. Mckeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pages 174–181, 1997.

[Hos05] A. Hoskinson. Creating the ultimate research assistant. *IEEE Computer*, 38(11):97–99, 2005.

[HPT+02] L. Hirschman, J.C. Park, J. Tsujii, L. Wong, and C. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

[JB06] N. Jindal and L. Bing. Identifying comparative sentences in text documents. In *29th SIGIR*, pages 244–251, 2006.

[JCD04] R. Jalam, J-H. Chauchat, and J. Dumais. Automatic recognition of keywords using n-grams. In *16th Symposium of IASC (COMPSTAT 04)*, pages 1245–1254, 2004.

[JSFT07] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WEBKDD/SNAKDD*, pages 56–66, 2007.

[KM01] A. Kohrs and B. Merialdo. Improving collaborative filtering for new-users by smart object selection. In *Oral presentation*, 2001.

[KMMR04] J. Kamps, M. Marx, R.J. Mokken, and M. De Rijke. Using wordnet to measure semantic orientations of adjectives. In *4th LREC*, pages 1115–1118, 2004.

[Koz93] H. Kozima. Text segmentation based on similarity between words. In *31st ACL*, 1993.

[KP04] A. Kao and S. Poteet. Report on kdd conference 2004 panel discussion - can natural language processing help text mining? *SIGKDD Explorations*, 6(2):132–133, 2004.

[KP06] A. Kao and S. Poteet. Text mining and natural language processing - introduction for the special issue. *SIGKDD Explorations*, 7(1):1–2, 2006.

[KPKF01] A. Kehagias, V. Petridis, V.G. Kaburlasos, and P. Fragkou. A comparison of word- and sense-based text categorization using several classification algorithms. *Journal of Intelligent Information Systems*, 21(3):227–247, 2001.

[Kri75] K. Krippendorff. *Information Theory. Structural Models for Qualitative Data.* Sage Publications, 1975.

[KU96] K. Kageura and B. Umino. Methods of automatic term recognition. *Technology Journal*, 3(2):259–289, 1996.

[Lew92] D.D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *SIGIR*, pages 37–50, 1992.

[Lin98] D. Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL*, 1998.

[Liu07] B. Liu. *Web Data Mining - Exploring Hyperlinks, Contents and Usage Data.* Springer, 2007.

[MB06] R.J. Mooney and R. Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explorations*, 2006.

[MBC$^+$05] D. Metzler, Y. Bernstein, W.B. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. In *CIKM*, pages 517–524, 2005.

[MBSC97] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *5th International*

Conference "Recherche d' Information Assistee par Ordinateur" (RIAO), pages 200–214, 1997.

[MC91] G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[McC05] A. McCallum. Information extraction: Distilling structured data from unstructured text. *ACM Queue*, 3(9), 2005.

[MCD08] S. Maurel, P. Curtoni, and L. Dini. L'analyse des sentiments dans les forums. In *Atelier FOuille des Données d'OPinions (FODOP 08), en conjonction avec le 26ème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 08)*, 2008.

[MG98] D. Mladenic and M. Grobelnik. Word sequences as features in text-learning. In *7th Electrotechnical and Computer Science Conference*, pages 145–148, 1998.

[Mob07] B. Mobasher. Data mining for personalization. *Adaptive Web: Methods and Strategies of Web Personalization, LNCS*, 4321:90–135, 2007.

[MS99] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing.* The MIT Press, 1999.

[NA06] G. Nenadic and S. Ananiadou. Mining semantically related terms from biomedical literature. *ACM TALIP Special Issue on Text Mining and Management in Biomedicine*, 5(1):22–43, 2006.

[NG00] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *8th CIKM*, pages 86–93, 2000.

[PB97] M. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27:313–331, 1997.

[PHLG00] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 473–480, 2000.

[PLV02] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[PRDP08] M. Plantié, M. Roche, G. Dray, and P. Poncelet. Is a voting approach accurate for opinion mining? In *10th International Conference on Datawarehousing and Knowledge Discovery*, 2008.

[RAC+02] A. M. Rashid, I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan, and J. Riedl. Getting to know you: Learning new user preferences in recommender systems. In *IUI*, 2002.

[RB99] M. Rajman and R. Besançon. Stochastic distributional models for textual information retrieval. In *9th ASMDA*, pages 80–85, 1999.

[Res95] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *14th IJCAI*, pages 448–453, 1995.

[Res99] P. Resnik. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[Ric79] E. Rich. User modeling via stereotypes. *Cognitive Science*, 3:329–354, 1979.

[Ril95] E. Riloff. Little words can make a big difference for text classification. In *18th SIGIR*, pages 130–136, 1995.

[RIS+94] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proc. of ACM Computer Supported Cooperative Work*, pages 175–186, 1994.

[Rob77] S.E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33:294–304, 1977.

[Sal88] G. Salton. Syntactic approaches to automatic book indexing. In *26th ACL*, pages 120–138, 1988.

[SAMK05] I. Spasic, S. Ananiadou, J. Mcnaught, and A. Kumar. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251, 2005.

[SAN07] A. Stavrianou, P. Andritsos, and N. Nicoloyannis. Overview and semantic issues of text mining. *SIGMOD Record*, 36(3):23–34, 2007.

[Sap21] E. Sapir. *Language: an introduction to the study of speech.* HARCOURT BRACE and CO., 1921.

[SC08] A. Stavrianou and J-H. Chauchat. Opinion mining issues and agreement identification in forum texts. In *Atelier FOuille des Données d'OPinions (FODOP 08), en conjonction avec le 26ème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 08)*, pages 51–58, 2008.

[Sch94] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference of new methods in language processing*, 1994.

[Sch99] R.E. Schapire. A brief introduction to boosting. In *16th IJCAI*, pages 1401–1405, 1999.

[Seb02] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[Seb06] F. Sebastiani. Classification of text, automatic. *The Encyclopedia of Language and Linguistics*, 14:457–462, 2006.

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[SL68] G. Salton and M.E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, 1968.

[SM95] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating "word of mouth". In *CHI-95*, 1995.

[SS94] D.R. Swanson and N.R. Smalheiser. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neuroscience Research Communications*, 15(1):1–9, 1994.

[SS97] D.R. Swanson and N.R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183–203, 1997.

[STE07] J. Scripps, P-N. Tan, and A-H. Esfahanian. Node roles and community structure in networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 26–35. ACM, 2007.

[SVC09] A. Stavrianou, J. Velcin, and J-H. Chauchat. Definition and measures of an opinion model for mining forums. In *2009 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 09)*, 2009.

[SVH04] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *16th ECAI*, pages 1089–1090, 2004.

[SWY75] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[TL03] P.D. Turney and M.L. Littman. Measuring praise and criticism: inference of semantic orientation from association. *ACM TOIS*, 21(4):315–346, 2003.

[TMR09] S. Trausan-Matu and T. Rebedea. *Polyphonic Inter-Animation of Voices in VMT.* Studying Virtual Math Teams, Computer-Supported Collaborative Learning Series 11, G.G.Stahl, Springer US, 2009.

[Tur02] P.D. Turney. Thumbs up or down? semantic orientation applied to unsupervised classification of reviews. In *ACL-2002*, pages 417–424, 2002.

[VG07] J. Velcin and J-G. Ganascia. Topic extraction with agape. In *Proc. of the International Conference on Advanced Data Mining and Applications*, 2007.

[vR79] C.J. van Rijsbergen. *Information Retrieval.* 2nd edition, Butterworths, London, 1979.

[VVR+05] G. Varelas, E. Voutsakis, P. Raftopoulou, E. Petrakis, and E.E. Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *7th WIDM*, pages 10–16, 2005.

[WBMT99] I.H. Witten, Z. Bray, M. Mahoui, and B. Teahan. Text mining: a new frontier for lossless compression. In *DCC*, pages 198–207, 1999.

[Wie00] J.M. Wiebe. Learning subjective adjectives from corpora. In *AAAI-2000*, 2000.

[Yar95] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd ACL*, pages 189–196, 1995.

[YHM03] A.S. Yeh, L. Hirschman, and A.A. Morgan. Evaluation of text data mining for database curation: lessons learned from the kdd challenge cup. *Bioinformatics*, 19(1):i331–i339, 2003.

[YP97] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *14th ICML*, pages 412–420, 1997.

[ZAA07] J. Zhang, M.S. Ackerman, and L. Adamic. Expertise networks in online communities: Structure and algorithms. In *Proc. of the 16th International conference on World Wide Web*, pages 221–230, 2007.

[Zai98] O.R. Zaiane. From resource discovery to knowledge discovery on the internet. In *Technical Report TR 1998-13, Simon Fraser University*, 1998.

[ZCT02] Y. Zhang, J. Callan, and T.Minka. Novelty and redundancy detection in adaptive filtering. In *SIGIR-02*, 2002.

Appendix A

Measures based on the Post-Reply Opinion Graph

In this Appendix, we present a summary of the measures that are based on the Post-Reply Opinion Graph.

Table A.1: Structure-oriented measures for a discussion represented by a PROG graph G.

Description	Formula
Root postings	$root(G) = \{v \in V : outDegree(v) = 0\}$
Root postings for a topic T_i	$root(G) = \{v \in V : v \in initialVertex(G, T_i)\}$
Leaf postings	$leaf(G) = \{v \in V : inDegree(v) = 0\}$
Leaf postings for a topic T_i	$leaf(G) = \{v \in V : v \in finalVertex(G, T_i)\}$
Popularity of a posting v	$inDegree(v) =\mid \{v' \in V : (v', v) \in E\} \mid$ $inDegreeExtra(v) =$ $\mid inVertices(v) \cup \ (\cup inVerticesExtra(i)) \mid$ $inDegreeDesc(v) =$ $\mid inVerticesExtra(v) \cup inVerticesExtra(v') \mid$, where $v' \in (descendants(v) \cap root(G)) \cap msgs(T)$
Order of the graph G	$order(G) =\mid V \mid$ $orderThread(G) =\mid \{G_{thr} \in G\} \mid$ $orderChain(G) =\mid \{G_c \in G\} \mid$
Order of a discussion thread G_{thr}	$orderOfThread(G_{thr}) =\mid V_{thr} \mid$ $orderThrChain(G_{thr}) =\mid \{G_c \in G\} \mid$
Order of a discussion chain G_c	$orderOfChain(G_c) =\mid V_c \mid$

Table A.2: Opinion-oriented measures for a discussion represented by a PROG graph G.

Description	Formula
Opinion status of a user u	$avgOpFromUsr(G_c, u) = \frac{\sum_i op_{v_i}}{\lvert msgs(u) \cap V_c \rvert}$ $avgOpFromUsr(G_{thr}, u) = \frac{\sum_i op_{v_i}}{\lvert msgs(u) \cap V_{thr} \rvert}$ $avgOpFromUsr(u) = \frac{\sum_i op_{v_i}}{\lvert msgs(u) \rvert}$
Opinion reactions towards a user u	$avgOpToUsr(G_c, u) = \frac{\sum_i op_{v'_i}}{\sum \lvert inVerticesV(v) \rvert}$ $avgOpToUsr(G_{thr}, u) = \frac{\sum_i op_{v'_i}}{\sum \lvert inVertices(v) \rvert}$ $avgOpToUsr(u) = \frac{\sum_i op_{v'_i}}{\sum_j inDegree(v_j)}$
Opinion reactions towards a posting v	$reply(v, r) = \mid \{v' \in inVertices(v), op_{v'} = r\} \mid$ $avgMsgOpinion(v) = \frac{\sum_i op_{v'_i}}{inDegree(v)}$ $H(v) = -\sum_{r=n,o,p}(\frac{reply(v,r)}{inDegree(v)} \log \frac{reply(v,r)}{inDegree(v)})$
Opinion information of a discussion chain G_c	$H(G_c) = -\sum_{r=n,o,p}(\frac{verticesCh(G_c,r)}{\lvert E_c \rvert} \log \frac{verticesCh(G_c,r)}{\lvert E_c \rvert})$

Table A.3: Time-oriented measures for a discussion represented by a PROG graph G.

Description	Formula
Duration of a discussion	$duration(G) = tm_{v'} - tm_v$, $tm_{v'} = argmax_{i \in V} tm_i$, $tm_v = argmin_{i \in V} tm_i$
Time-comparison of two postings	$compareTime(v, v') = \frac{tm_v - tm_{v'}}{\lvert tm_v - tm_{v'} \rvert}$
Ancestors of a posting beyond structure	$ancestors(v) = \{v' \in V : compareTime(v, v') = 1\}$

Table A.4: Topic-oriented Measures for a discussion represented by a PROG graph G.

Description	Formula
Postings that belong to a topic T	$msgs(T) = \{v \in V : topic(v) = T\}$
Participation of a user u in a topic T	$userParticip(u,T) = \frac{\lvert msgs(u) \cap msgs(T) \rvert}{\lvert msgs(u) \rvert}$
Opinion evolution of a user u in a topic T	$opEvolution(u,T) = \frac{\sum_i op_{v_i}}{\lvert msgs(u) \cap msgs(T) \rvert}$ or $opEvolution(u,T,tm_v,tm_{v'}) = \mid op_v - op_{v'} \mid$
Opinion expressed for a topic T	$avgOpTopic(T) = \frac{\sum_i op_{v_i}}{\lvert msgs(T) \rvert}$ $avgOpTopic(G_c,T) = \frac{\sum_i op_{v_i}}{\lvert msgs(T) \cap V_c \rvert}$ $avgOpTopic(G_{thr},T) = \frac{\sum_i op_{v_i}}{\lvert msgs(T) \cap V_{thr} \rvert}$
Topic Popularity	$topicPop(G_c,T) = \frac{msgs(T) \cap V_c}{\lvert V_c \rvert}$ $topicPop(G_{thr},T) = \frac{msgs(T) \cap V_{thr}}{\lvert V_{thr} \rvert}$ $topicPop(G,T) = \frac{msgs(T)}{\lvert V \rvert}$
Ancestors of a posting by topic	$ancestors(v,T) = \{v' \in V : tm_{v'} < tm_v, topic(v') = topic(v)\}$

Table A.5: User-oriented Measures for a discussion represented by a PROG graph G.

Description	Formula
Users who initiate a sub-discussion	$usersStartDisc(G) = \{u_v : v \in root(G)\}$
Users who end a sub-discussion	$usersEndDisc(G) = \{u_v : v \in leaf(G)\}$

Appendix B

The Online Discussion used in Chapter 3.

This is a real web discussion in French taken from the site http://www.liberation.fr/ regarding the open-space offices. Each line contains the message id as this is generated by our application and the content of the message as it appears on the web site.

Title : Pour ou contre les bureaux open space.

279 & Avantage : Une certaine convivialité. On connait ses collègues, on peut plus facilement créer des affinités, ça humanise les rapports humains. Inconvénient, effectivement, on n'a aucun espace à soi, la vie privée est difficile à préservée. Impossible (en tout cas très difficile), de s'isoler totalement pour se concentrer sur son travail, bruit permanent, bavardages, commérages, engueulades (enfin, la vie de groupe quoi !). Dans l'Open Space, parfois, j'ai l'impression d'être un lofteur filmé en permanence ! Mais j'aurais du mal à me retrouver dans un bureau seul, ou avec une seule personne ! Si on ne la supporte pas, ça peut vite devenir l'enfer. Et les petits bureaux coupent la communication entre collègues... Dernier inconvénient de l'open space : Quand on veut draguer une collègue, on se fait gauler direct ! Et si on se prend une veste, c'est encore plus vexant ! ;-)

280 & J'ai eu l'occasion de travailler dans des open-spaces et dans dans un bureau plus standard; Etant marseillais, je suis certainement plus extraverti que d'autres, mais je préfère largement un open-space, à certaines conditions: - Avoir la chance d'être entouré de collègues sympas (je sais, ce n'est pas toujours le cas) - Avoir un minimum d'espace vital (je me trouve à 4 mètres de mon voisin le plus proche). - Que les gens évitent d'utiliser la fonction "main-libres" de leur téléphone; le bruit devient vite gênant: la solution, un micro-casque... Le travail en bureau fermé nous isole très vite, et rompt le tissus social d'une entreprise; j'ai été 14 syndicaliste, et ce tissus, j'y tiens particulièrement. Pour

moi, donc, les bureaux paysagés le sont pas spécialement synonymes de flicage; mais tout dépend des finalités de l'entreprise dans laquelle on travailleĚ

281 & Vous faites vraiment... suer avec votre anti anglicisme à tout va! Proposez une alternative au lieu de jouer l'opposition stérile à tout va! Il se trouve que les franchouilles sont un peu lents à la détente pour lancer de nouveaux termes et que de toute facon, le monde entier se fiche de ce que dit la France depuis au moins 200ans... Courriels, pourriels, les clavier azerty. de l'exception au ridicule, il n y a parfois qu un pas

282 & le clavier AZERTY. L'ordre des lettres sur un clavier provient de savantes études sur leur fréquence d'utilisation dans une langue donnée et sur l'optimisation mécanique des anciennes machines écrire (pour éviter que les tiges supportant les caractères s'empilent trop souvent les unes sur les autres pendant la frappe). s'y ajoute les particularité locales comme les lettre accentuées, les symboles monétaires... Chaque langue à SON clavier, le clavier allemand possède son SS et nous on a les é,è,à,ù etc etc

283 & Accents a part, les etudes pour la creation du clavier Azerty n'ont pas ete effectuees par des "savants" tres brillants! Exemple avec le point, une des touches les plus utilisees qui ne s'accede pas directement mais avec un "shift x". Une mise a jour serait souhaitable a mon avis.

284 & Le clavier azerty n'a pas été étudié à partir de la fréquence d'utilisation des lettres ! il a été étudi à partir de la probabilité que 2 marteaux se coincent... je m'explique : à l'origine, c'étaient des machines à écrire et lorsqu'on appuyait sur une touche, un marteau était projeté sur un ruban encré, qui imprimait une lettre sur la feuille. Les marteaux étaient rangés les uns à coté des autres, de gauche à droite, comme le clavier. Il fallait éviter que 2 marteaux ne risquent de se croiser et s'entrechoquer en étant physiquement trop proches. Donc on a mis proches des lettres qui ne se trouvent pas à coté dans des mots. Le clavier vraiment pratique et économisant les muscles, c'est le dvorak, mais personne ne l'utilise parce que tt le monde s'est habitué à l'azerty (moi le 1er ;-))

285 & J'ai appris quelquechose aujourd'hui! Et bien je dis raison de plus pour mettre a jour ce type de clavier créé pour une technologie archaique! :)

286 & azerty, dvorak ou quoi que ce soit, le problême, c'est que la plupart des gens qui utilisent un clavier n'ont jamais appris à taper à dix doigts.

287 & Ce que c'est chiant ce racisme anti-français:"ah, en France, on est des gros nuls, ah en France, ça "bouge pas", ah en France on bosse moins (désinformation, il est prouvé que les français bossent plus que les anglais ou les allemands mais les préjugés ont la vie dure et ça arrange la droite de nous traiter de feignasses, alors...), ah en France on est arrogant...mais tirez-vous aux Etats-unis ou au Japon, vous verrez si c'est si génial!! quand vous aurez un cancer pas d'assurance maladie, quand vous aurez un patron qui a tous les droits, quand vous aurez 2 semaines de vacances etc...marre de ce poujadisme à la con, en plus, ces français qui méprisent les autres français et le fonctionnement de leur pays sont

souvent les 1ers à profiter du système et, évidemment, NE SE METTENT PAS DANS LE LOT, eux, ils sont différents, ils parlent des AUTRES français, bien sûr...

288 & je suis d'accord, allez bosser à NY ou Hong Kong, vous verrez ce que c'est l'esclavage moderne proné par l'ultra libéralisme. Les anglos saxons font peut être plus d'argent, au prix d'une souffrance humaine qui se tait.

289 & T'as raison! En France il n'y a pas de pognon, mais au moins personne ne souffre!

290 & Anneso, Tout a fait d'accord avec toi, et je suis bien place pour le savoir, car je vis aux USA depuis 9 ans et travaille dans une des plus grosses boites americaines. Je suis parti aux USA non pas parce que je voulais me la jouer, comme ces cadres francais qui essayent de singer les anglosaxons et truffent leur conversation de mots anglais, ce qui m'a toujours fait sourir (je ne veux pas etre blessant!), mais sur ultimatum de ma copine americaine (devenue mon epouse legitime). J'ai mon propre bureau personnel et je m'en trouve tres bien. Cela me permet d'etre efficace, et si j'ai besoin de communiquer, il y a les emails, le telephone et tout simplement une bonne discussion face a face, et je communique avec enormement de monde que ce soit sur mon lieu de travail ou ailleurs autour du monde (on n'a pas encore resolu le probleme du decalage horaire). Pour le reste, ceux qui n'ont pas de bureau personnel ont des "cubicles", espaces de travail personnels avec des cloisons semi-fermees qui permet d'eviter d'etre au vu de tout le monde. Comme ils disent ici, "the grass is always greener on the other side". Et les Francais qui passent leur temps a critiquer leur pays feraient bien d'aller tater du modele anglo-saxon et d'emigrer comme le pequin moyen aux USA. Je ne parle bien entendu pas des expatries Francais qui apres un an ou deux aux USA reviennent au pays en se la jouant a donf, alors qu'ils sont restes dans le cocon de leur boite francaise avec tous les avantages conserves bien entendu. Moi, je n'ai qu'une envie: revenir au pays souffler un peu!

291 & quel sombre crétin pourquoi devrai-ton se taire? faire des propositions? bin retour a ce qui a fait ses preuves 'pourquoi faire comme le monde entier quand c'est stupide ou pas dans nos moeurs pourquoi voudrais-tu imposer ta vision aux autres? hummm ca m'interpelle toujours les personnes comme toi...dans le fond tu es assez oppressant et puis c'est vrai qu'en france on est plus con que les autres hein y'a qu'a regarder notre productivité à nous les feignants de francais le mieux en fait c vivre sa vie sans pourrir celle des autres ;)

292 & Ce qui me fait marrer, c'est l'utilisation de l'anglais pour présenter sous un jour agréable et moderne des réalités parfois sordides. Tiens, pourquoi on n'appellerait pas les dortoirs pour SDF des chambres open space ?

293 & L'utilisation d'anglicisme, on le sait depuis longtemps, sert à masquer le manque de fond d'un discours. Prenez un discours farci d'anglicisme ou autres balivernes pseudo technique et traduisez le en français. Il ne reste plus grand chose.

294 & Oui Je suis contre, par experience! dans l open space tout est calculé pour le

controle. On est a moins d un metre de son collegue, impossible de passer un coup de fil, vos collegues savent tout sur votre vie privé, donc elle s efface et il ne reste que le travail. Est ce la normalite? Est ce ecrit dans la Bible ou dans l evolution qu on doit faire 8 a 10heures pas jour sans meme pouvoir penser a ses enfants, sa famille, le pain pour le soir... C est de l esclavagisme moderne. 96% des francais sont aujourd hui employés; contre 16% il y a 50 ans, il n y a plus d esprit d initiative, que des petits robots tristes et stressés qui essaient de caser un peu de vie dans leur RTT... qu on va supprimer.

295 & “Est ce ecrit dans la Bible ou dans l evolution qu on doit faire 8 a 10heures pas jour sans meme pouvoir penser a ses enfants, sa famille, le pain pour le soir... ” Ben c'est plutôt le contraire qu'on lit ici ou là dans le forum et qu'on vit au quotidien dans le travail. En fait , c'est carrément l'envahissement, l'interpenétration de la sphere professionnelle par la vie intime , familiale ou privée de chacun, les petites histoires persos au telephone ou avec la voisine de bureau, les petites ou grandes miseres de l'existence, l'expression publique de ses etats intimes pas forcement en rapport avec le boulot, bonne humeur, deprime, anxiete, joie, amour, la montée en puissance du jeje - moimoi, “ regardez- ecoutez, vous avez vu les gens comment je suis super bien (sous vos applaudissement) ou super mal (sous votre compréhensive commisseration obligatoire) dans ma vie à moi que j'ai ”. Et le droit à l'indifférence ? à l'etanchéité entre la vie perso et la vie professionnelle ? le droit de ceux (rare) qui sont envahi 8 heures par jours par la vie intime des autres (nombreux et bruyants) et qui n'ont ni envie de raconter leurs histoires ni envie de se fader celles des autres. Juste de faire leur taff et basta? Et si chacun se diciplinait un peu pour pas bouffer l'espace open ou fermé de son voisin de boulot ? Et tiens une idée absurde comme ça en passant, si on venait au boulot pour bosser, faire ses heures , pas pour raconter sa “life” et en faire profiter tout le monde...?

296 & quel cynisme et quelle froideur; j'aimerais pas me retrouver dans un open space avec vous; ça doit pas rigoler tous les jours...mais c'est vrai, on est au boulot, on est pas là pour rigoler, sorry!

297 & J'en ai connu à mon travail des comme toi qui ne supportaient pas qu'on ait une vie en dehors du travail parce que probablement ils n'en ont pas, de vrais schizophrènes qui ont créé un tel fossé entre leur personnalité privée et publique que çà fait peur, en général ce sont les plus arrivistes et les plus léche-Q qui sont comme çà, et qui tirent des tronches de 10 m en plus! Moi j'ai signé pour travailler dans un bureau ok, pas pour être moine dans un monastére!

298 & Moi, j'ai signé pour un boulot et un salaire. La boite où je bosse c'est pas ma famille, et les gens qui y bossent c'est pas mes potes, c'est des collégues. Ma vie perso, je la vis ailleurs que dans mon entreprise: espace open -cool -on-est-tous-des-potos ou pas J'ai pas signé et je ne suis pas payé pour faire la clape de tous les petits malins qui passent leur temps à expliquer comment ils sont debordés et stressés en faisant leur course sur internet pendant que les studieux essayent de se concentrer pour bosser , se coltinant un peu plus que leur part , même si ils ne la souhaitent pas. Parce qu'il ya un principe de base dans le taff: le taff va au taff .L Les chefs de service sont moins cons qu'il n'y parait,

c'est pas sur le petit malin qui fait le spectacle dans l'open espace qu'il fout la pression, il sait qu'il aura peanuts comme rendu de boulot en plus , mais sur celui qui bosse en silence , là où il y a des chances d'avoir du retour.... Pour faire la mesure, ces collégues là , non seulement ils t'emmerdent quand tu bosses, n' ont aucun scrupule à laisser le ollectif assurer pour eux le boulot qu'ils ne font pas mais en plus ils se permettent de venir baver sur le boulot des autres.... C'est les mêmes qui sont persuadés qu'occuper un poste est synonyme de travailler et qu'en plus il faudrait leur derouler un tapis rouge pour accepter d'être là le matin , tellement ils sont plus geniaux que les autres.

299 & Helas, je confirme..En tant qu'ancien membre de l'encadrement, c'est tout a fait vrai que l'important lorsqu'on dirige une equipe est d'en connaitre ses membres, de savoir lesquel sont fiables et de confiance et qui sont les rigolos. Et bien entendu, lorsqu'on a des resultats a obtenir, on sait sur qui on peut compter...Pour le reste, l'equite n'est pas une notion tres importante en management, surtout que les rigolos sont en general retors et sont promptes a vous creez des problemes. Alors, on evite d'avoir des ennuis avec la DRH, et on charge toujours les memes! Je le deplore, et c'est d'ailleurs pour cela que je ne fais plus d'encadrement, trop chiant et frustrant lorsqu'on a un fonds d'honnete homme.

300 & je croyais être le seul à penser cela mais je vois que vous êtes comme moi : et si les gens allaient au boulot pour bosser au lieu de raconter leurs vies ?

301 & difficile de bosser dans un open-space, où la plupart du temps, on fait tout sauf bosser, avec sous les yeux et dans les mains, un outil pour ne pas bosser non plus, msn et autres (je sais dire les yeux fermés si quelqu'un tchate ou bosse, à sa façon de taper sur le clavier), et ça rend fou ... je ne bosse pas en dilettante, je suis professionnel, et ça empêche pas de péter un plomb et mettre un "big bisous" au taquet sur les enceintes histoire de décompresser, mais rester au milieu de cette foule molle du genou avec chaque jour ces histoires insipides dans lesquelles "ohlala c'est pas facile", ça use et ça impacte sur ma productivité, ce qui m'ennuie profondément, n'étant pas là uniquement pour chopper un salaire, mais faire un taf qui me gratifie. Pour avoir bossé dans une start-up, où l'open-space fait partie de la religion, je sais bien que c'est un outil de branleur, que cette forme de bureau est sortie de 'limagination d'un mec qui avait envie de ne rien foutre, tout en se faisant passer pour un vrai bosseur : c'est la vertu de l'open space, on branle presque rien, on met deux ou trois fois le temps nécessaire à éxécuter une tache, et résultat, on passe pour un acharné qui a réussi un exploit ... c'est garantit. Si vous supportez cette ambiance, vous êtes soit sourd, soit doué d'une capacité de concentration qui vous ouvre les portes des championnats d'échecs mondiaux, soit vous êtes un de ces mollusques. Le tabac nuit gravement à votre entourage, l'open-space aussi !

302 & oui, qu'à moitié. En effet, il ne faut se soumettre ni à l'une ni à l'autre des extrêmes Lali, ton patron ne t'envoudra pas si tu discutes d'un match de foot, ou de l'éléction d'Obama 5min avant de te mettre au travail (Surtout si le travail d'équipe est nécessaire, les anglais, peut-être plus cons ... font même des scéances de 'TeamBuilding' ==> ça m'étonne toujours en France de voir le nombre de ceux qui se défilent quand on leur propose d'aller boire un verre avec l'équipe après le boulot !) En tout cas, idée à

cultiver en France plus que toute autre : Il y a un juste milieu à tout, ça suffit du système binaire du tout ou rien, c frustrant à la fin, il faudrait pouvoir temporiser. Cheers

303 & Ah les "team building sessions"! ca me fait toujours rire, moi qui bosse aux USA. Les Americains ne sont pas aussi cons que les Francais veulent bien le croire. Ils sont anglo-saxons, ce qui signifie qu'ils n'exprimeront jamais ce qu'ils pensent reellement. C'est considere comme impoli. Mais ils savent eux qu'ils n'ont pas de potes au boulot, uniquement des collegues avec qui ils sont en competition permanente, vu qu'on est sans cesse evaluee par nos superieurs, et qu'a la fin de l'annee le bonheur des uns (augmentations, primes, etc..) fait le malheur des autres (0% d'augmentation, retrogadation, avertissement, licenciement). Cela n'empeche pas de travailler ensemble en bon ordre et sans trop de probleme, car chacun accepte ce systeme, n'ayant pas d'autre choix.

304 & On ne développe pas la rentabilité d'une activité en rognant sur la surface, on ne change pas les références culturelles sur une décision comptable ! Et si l'espace dans lequel on vit correspondait proportionnellement au volume que l'on s'accorde pour penser, il faudrait tenir compte du nombre de gens avec lesquels on partage cet espace : comparons hauteur de plafond et pouvoir ; profondeur de champ visuel et analyse profonde (certes, il existe aussi le "crétin des alpes") ; Volume de l'espace de vie et pouvoir de décision. Au USA nos WC passent pour des placards à Balais. Ici nos HLM aux normes standard (disent-ils) fabriquent de la délinquance ou du misérabilisme à plein tuyau ... Vivons étroit pour penser droit ?

305 & AlcideH, vous me faites tousser... moi qui vit aux USA depuis 9 ans. Je regrette nos WC "placars a balais": au moins c'est intime, ferme, et les odeurs restent dans l'endroit auquelles elles appartiennent. Aux USA, les WC sont dans les salles de bain. Agreable de prendre sa douche ou son bain lorsque les WC viennent d'etre utilises par ma femme lorsqu'elle a une urgence! (a moins d'avoir le nez bouche). quant a nos HLM, y ayant moi meme vecu enfant, votre reflexion me fait rire. Ce ne sont pas les HLM qui font les delinquants, c'est le chomage, l'echec scolaire, des parents absents ou depasses, l'absence d'espoir et de futur. L'HLM ne m'a pas empeche de faire des etudes superieures, de devenir cadre, de partir aux USA, et de survivre plutot bien (jusqu'a present bien sur!) . Les Americains sont tres mal place pour donner des lecons quant a la fabrication de delinquants. Ou plutot si, ils sont plutot bons dans le domaine car la creation de delinquance aux USA, c'est une industrie nationale car ca cree du business (prisons privees ou les detenus doivent travailler pour des peanuts - c'est une activite tres juteuse et cotee en bourse, industrie de la securite qui ne s'est jamais aussi bien portee, avec la vente d'armes et equipements de securite a tout le monde, etc...). Tiens, cela me fait penser... je n'ai toujours pas exercer mon droit d'acquerir mon flingue perso (pour 300 euros, je pourrais enfin jouer les Clint Estwood). Je suis helas reste trop Francais!

306 & Je suis pour l'open space,mais un par personne. On appelle ca aussi les Landscaped Offices (bureaux paysages).

307 & En fait, les aménagement de l'espace qui nous conviennent reposent sur deux

choses : des constructions mentales inconscientes et des représentations culturelles dont l'origine a été oubliée. Lorsque l'on présente un plan de bureaux qui doivent comporter des zones à accès restreint deux modèles fonctionnent pour le client : celui du temple grec (le saint des saints est au fond) et celui du temple de Salomon (le saint des saints est au centre). Les agencements de bureaux, quelle que soit la disposition ont toujours une fonction de contrôle social. Le bureau "paysager" n'offre qu'un seul avantage : il permet théoriquement de moduler plus facilement la taille des compartiments. Personne n'utilise ou presque cette propriété qui est facilement limitée par la disposition des points d'accès aux réseaux et à l'énergie. La plupart des entreprises utilisent en fait un mélange d'espaces ouverts et d'espaces clos. Il existe en fait des bureaux paysagers bien aménagés et des bureaux paysagers mal fichus, voir glauques. Prétendre que cette disposition est récente est assez burlesque. Certaines professions l'ont utilisée depuis longtemps pour des raisons pratiques : favoriser la pénétration de la lumière notamment. De la même manière, il existe plusieurs manières de gérer un espace ouvert ou fermé. Dans certaines entreprises, certains salariés ou sous-traitants n'ont pas de place réellement attribuée et les occupent en fonction de la disponibilité. Ceci interdit aux employés de s'approprier l'espace et de le décorer à leur manière, facilité l'activité des personnes chargées du nettoyage des lieux et prévient aussi les vols. En résumé, la question n'est pas pour ou contre telle ou telle forme d'aménagement, mais quel rapport l'entreprise veut-elle entretenir avec ses employés au travers du rapport avec l'espace ?

308 & C'est curieux pour ma part j'ai découvert les bureaux sur plateau avec des cloisons à mi-hauteur pour les chefs de service en.....1972 !!! alors ce n'est pas vraiment nouveau . J'ajouterai que les gens qui travaillent dans des ateliers ,dans des grandes surfaces,sur des chantiers etc sont eux aussi dans des open spaces mais qu'il n'en est pas fait grand cas n'est ce pas ?

309 & Faudrait peut-être pas comparer une personne qui range les rayons qui ne demande pas un effort intellectuel important à celui qui bosse dans un bureau et devant produire un effort intellectuel important donc un niveau de concentration plus important nécessitant un isolement plus important. Merci de comparer ce qui est comparable !

310 & M'sieur se la péte ....

311 & en identifiant travail de bureau en travail intellectuel. Tu te montes pas la tête tout seul? Le travail de bureau chauffé l'hiver et climatisé l'été avec des horaires fixes.

312 & pour l'avoir connu, contre définitivement ! encore 1 astuce pour augmenter la productivité! en faisant fi de toute intimité, tout le monde surveille tout le monde....franchement "les temps modernes" de Chaplin sont toujours d'actualité....et ce n'est pas une franche avancée de la liberté humaine bien au contraire. a quand la puce dans le bras, déclinant votre identité vos jours de congé, de RTT, et de travail effectif...avec décharge électrique au moindre relâchement d'attention ! Il faut refuser ça et vite !

313 & Ca n'engage que les managers qui veulent y croire... Euh qui sont dans des

bureaux personnels.

314 & en fait tout le monde a voulu faire comme chez Google, Apple, Microsoft... à la différence que ceux-ci n'engagent que des génies (dira-t-on des geeks) avec un profil pour lesquels un quota de distraction, de relation sociale, de développement de soi, de soutien, est indispensable. Bref l'openspace était à la base pensé pour l'épanouissement personnel à but de productivité avec des mesures généreuses (temps libre pour des projets mais un cadre bien défini pour l'évaluation etc.) Dans la plupart des boîtes on applique l'openspace à des tâcherons (désolé si j'exagère) pour lesquels le modèle n'est pas pertinent. Après en fonction des personnes et de l'ambiance, ça peut être le paradis comme l'enfer (nous ne sommes que des hommes). La différence c'est que par exemple si chez Google l'openspace devenait nuisible, il ne se passerait pas un mois avant que l'on remonte des cloisons.

315 & ce qu'il faut interdire, c'est les 3x8 !

316 & J'ai oublié de mentionner le "chef" dans ce que j'appelais son aquarium. J'étais directement dans sa "ligne de mire", à force, nous ne pouvions plus nous voir en peinture. J'avais d'ailleurs complété le tableau en mettant sur mon armoire un poster avec des poissons exotiques et une bougie en forme e poisson, histoire de rire un peu (intérieurement).

317 & comme partager son bureau avec une végétarienne (merci les gaz) qui commente a voix haute ce qu'elle est en train de faire et qui vous tient au courant de toutes les étapes de sa ménopause! Au moins en open space il y ade la diversité....

318 & Ca confirme une tendance lourde: le pire dans le boulot, c'est pas le boulot, ni la configuration des lieux, c'est les gens avec qui on travaille.

319 & Le travailler ensemble, c'est l'enfer.

320 & L'avenir appartient à ceux qui bosseront de chez eux, à peine quatre heures par jour. C'est possible.

321 & Je travaille de chez moi mais pas 4h par jour plutôt 2x ou 3x 35h par semaine. Travailler de chez soi c'est aussi accepter de se soumettre à la dictature des objectifs qui s'accroîssent, car le patron part du principe qu'il n'y a pas de temps de transport donc plus de disponibilité pour atteindre des objectifs toujours plus élevés. Il faut faire face à des discours comme quoi l'on est privilégié, etc. donc l'on se sent presque obligé de faire profil bas et de tout accepter. Résultat : plus de limite entre travail et loisirs, puisque la vie toute entière devient consacrée au travail, et donc encore moins de vie privée que dans des horaires carrés passés en compagnie de ses collègues. Des avantages il y en a des inconvénients aussi et pas que celui-là. Par exemple, la dépense d'électricité qui explose. Et qui paie les notes de téléphone ? Pas toujours facile de négocier... Enfin si vous dites que vous travaillez tout le temps c'est difficile de vous croire car on ne peut vous surveiller. Quelque fois, la surveillance a du bon. Donc moi je dis que travailler de chez soi, ça ne doit pas être un rêve ou un idéal à atteindre, mais plutôt quelque chose de mûrement réfléchi,

et dont les avantages tant que les inconvénients doivent être bien pesés. En tout cas, ça ne peut convenir à tout le monde.

322 & vous avez mal géré le passage en télétravail, apparemment.

323 & n00w, respirez un bon coup. Le télétravail ne peut fonctionner qu'à partir du moment où TOUT le monde est content. Vous, le patron et les collègues. Je suis en télétravail complet depuis 3 ans (soyons honnête, ma boite est à 700 km d'ici). Vous ne pouvez pas tout exiger sans offrir de contrepartie. Pour ma part, je commence à 8 heures et plus à 9h30 (bye les embouteillages), je suis moins fatigué, moins stressé. Je mange ce que je veux et ne dépense pas des fortunes en sandwiches infects, je ne prends plus congé pour mes démarches administratives.Si ma fille est malade, pas besoin de garde. Alors oui, je consomme certainement plus de courant mais ma facture de carburant a drastiquement chuté. Je ne chauffe quasiment pas c'est vrai. Il est normal que votre employeur y trouve également son avantage. Je suis consultant sans aucun moyen de mesurer ma 'performance'. De ce fait, mon patron ne peut que compter sur ma bonne foi quant au travail effectué et aux heures prestées. De ce fait, je ne rechigne pas pour un coup de téléphone à partir de ma ligne fixe ou pour des frais excessifs. Je veux bien que l'on demande le remboursement de la dépense d'électricité, mais il faut d'abord quantifier cette dépense. Dans pas mal de cas, l'employeur participe ou rembourse l'abo internet. Sauvons le télétravail !

324 & Je suis interprète par téléphone. Mon boulot serait complètement rébarbatif si je ne pouvais pas le faire à la maison. Ma boite paie pour le téléphone et la connection internet. J'économise énormément en temps et argent, sans compter une réduction du stress. Plus de dépense pour des fringues professionelles, je m'habille en jean et t-shirt si ce n'est pas carrément pyjama ou le jogging toute la journée. J'ai mon chien et mon chat assis à coté de moi quand je travaille, c'est super! Je suis tout aussi efficace que si j'étais dans un centre d'appel et de bien meilleure humeur!

325 & j'ai travaillé à la Tour Montparnasse, dans ce genre de bureau. Nous étions 35 à l'agence comptable, du temps où les calculatrices étaient bruyantes ... dur dur de se concentrer ; j'ai fort apprécié quitter ce lieu, pourtant privilégié (époque de la canicule de 1976 - merci la clim) pour me retrouver dans un petit bureau individuel (même sans clim). Jamais je n'aurais souhaité retravailler ainsi dans le bruit et le manque "d'intimité"

326 & Il me semble que ce concept anglo saxon, a la pertinence de permettre l'humiliation, actes de contrition devant son supérieur, qui est l'âme même du management "a la française". Il suffit d'avoir travailler en open space en france pour comprendre que cela permet aussi d'attraper les personnes et en faire des exemples, car a défaut du son vous avez l'image. Et dans un système de management par soumission , pure produit des écoles, vous ne pouvez imaginez ceux a quoi j'ai assisté..

327 & je souscris tout à fait à ce propos. l'open space produit par nature un management qui ne peut que dériver. j'ai d'ailleurs planté ma dèm parce que ma boite me

proposais qu'une seule solution : bosser en open space en banlieue parisienne . j'y ai passé 2 jours et hasta la vista baby... après 20 ans de boite

328 & l'autre jour j'étais plié, j'ai fait un prout silencieux mais farci au milieu d'un groupe où je passais par hasard sans que l'on fasse attention à moi(promiscuité oblige) dans l'open et j'ai passé mon chemin, il y avait 2 nanas tops avec qq supermen du gratin. Il s'est passé un certain temps avant qu'une réaction de suspicion ne se généralise - j'en ris encore :)))

329 & Ben oui, cela fait assez marrer en fait. Tout le monde sait que ce n'est pas la panacée (bon, ce n'est pas la cata non plus, depuis que j'ai mis un cactus sur mon bureau, je vais beaucoup mieux) pourtant nos journalistes favoris nous publient régulièrement des études pour mieux nous le montrer. Dommage que nos patrons de plus en plus nombreux à adopter ce type de configuration ne soient pas convaincus. Il y a sans doute un bon compromis à trouver; un open space de plusieurs centaines de personnes c'est une grande scène de vie qui rassure un patron certes, un lieu serein de travail productif, moins sûr! Evidemment, le patron lui, il a un bureau parfaitement étanche au commun des mortels, bruit, odeur etc...mais il reste, je pense sincèrement persuadé (grâce aux economies substentielles de loyer) que ce type d'aménagement permet une meilleure communication et un meilleur management; si ça lui fait plaisir! Contrairement à lui, nous savons nous adapter, c'est d'ailleurs ce qui nous différencie principalement. Mais à décloisonner, il faut qu'il sache que tout se décloisonne. Au début le regard du voisin dérange, à la fin tu fais tes achats de Noël sur le net en plein après-midi sous le regard vaguement consentant de ton chef qui fait pareil, les réunions se font dans le couloir, les coups de fils dans les toilettes. Ton voisin déjeune sur son bureau sans complexe, les pauses café s'éternisent. Parfois les pots se font sous l'il blasé de personnes non invitées. Bientôt mon voisin prestataire viendra avec son lit de camp; du moment qu'il ne bouge pas trop la nuit la lumière restera éteinte et l'alarme ne se déclenchera pasĔ

330 & "Tout le monde sait que ce n'est pas la panacée (bon, ce n'est pas la cata non plus, depuis que j'ai mis un cactus sur mon bureau, je vais beaucoup mieux) pourtant nos journalistes favoris nous publient régulièrement des études pour mieux nous le montrer." et si vous visitez les locaux de libération rue Béranger : c'est un grand OPEN SPACE. on comprend tout.

331 & Prennons le problème à l'envers : depuis l'existance de l'usine, les ouvriers ont travaillé dans des espaces ouverts avec comme conséquence : Control mutuel Cout en metre caré par tete plus faible - Pas de confidentialité - Gènant pour la réflexion Pas difficile à comprendre, y compris par les directions. Les gens positionnés de nos jour dans les open space le sont car ils sont simplement des executants sans plus de responsabilité qu'un ouvrier comptable, ouvrier informaticien qui doivent appliquer des méthodes en évitant de mettre le grain de sable dans les rouages. Désolé les maillons mais toute notre génération de diplomés ne peut pas faire une génération entière avec uniquement des postes à responsabilités.

332 & ingenieur prestataire chez un grand operateur mobile et avant ca chez un fournisseur d acces en encore avant hotlineur je n ai rien connu d autre que les open space depuis mes stages d etudiant. plutot contre le principe, mais les collegues qui surveillent rien de plus facile que leur rendre la pareil donc le statut quo est de rigueur et personnelement mes chefs je les entend arrivé de loin. on en fait tout un drame. Je prefererais avoir une remuneration a la hauteur de mon travail (un internet touche entre 50% et 75% en net de plus que moi) qu avoir mon propre bureau.

333 & L'open space est un système de torture pour les salariés. Il faut démolir ces nouveau instruments de régression salariale.

334 & Je ne supporte plus l'open space. Il y a un effet cocktail party générateur de stress qui empêche de se concentrer. Et en plus il faut toujours adopter une positive attitude et un enthousiasme de façade. Cette dictature de bonheur est bien décrite dans le bouquin "l'open space m'a tuer" d'Alexandre des Isnards et Thomas Zuber. Je me suis bien reconnu dans certaines saynètes!

335 & Déjà qu'à 3 ds un bureau je trouve ca assez infernal... Actuellement stagiaire, je rêve déjà de ma propre entreprise comme d'un monde parfait avec des espaces fermés. Mon refuge au travail, ce sont les toilettes. C'est dire... Quand je pense que mon avanir proche se situe probablement dans un open space, ca me donne envie de devenir femme au foyer. Nan sérieux, ca me déprime que les modèles les plus cons et les moins appréciés soient toujours adoptés. Ca sonne un peu conservateur, mais on est bien passé du boeuf bouguignon au MacDo, parce que c'est plus fun. Lá c'est la même. Les salariés ont trouvé au début que c'était fun et on laissé faire les directions. Maintenant, on rame pour revenir en arrière. Enfin, dés que ma boîte avec des "espaces fermés" sera montée, je fais signe aux lecteurs de Libé ;)

336 & de tout coeur avec vous. Je déteste plus que tout les open spaces. Je rêve d'avoir de nouveau un bureau à partager avec 3-4 collègues grand maximum.

337 & Les WC open space arrivent et vont faire un malheur.

338 & En Chine, et c'est pas des bétises. J'ai même été obligé de l'expérimenter lors d'une "urgence". Je peux vous dire qu'on y évite le regard de ses voisins.

339 & Nous vos parents nous n'étions pas en (open... je ne sais même pas le dire mais c'était pas mieux : si la tranquilité je l'avais 2 dans un bureau et même en plus séparée par une demi cloison le rêve pourrait-on dire et bien non ! j'ai beaucoup souffert d'être seule sans compagnie face à mes chiffres et a mes responsabilités j'avais pour tout réconfort le patron qui passait sa tête de temps en temps et me demandait "ça va ?" j'aurais tellement aimé avoir de la compagnie... quand a travailler à la maison il y de nombreux frais supplémentaires en plus de la facture EDF ...

340 & Ca me fait hurler ces personnes qui disent qu'être seul(e) dans un bureau, c'est

triste... Vous n'avez pas de vie sociale en dehors du bureau ? C'est peut-être cela le fond du problème.

341 & Merci à Libé d'ouvrir ce genre de débat qui serait saugrenu dans la plupart des autres pays à commencer par la Belgique . A 50 ans comme juriste dans une banque belge, je vis mal l'open space mais je sais que c'est 10 x pire dans les entreprises asiatiques ou anglo-saxonnes. Ce n'est pas une question de respect ou de dignité de l'individu puisque ces valeurs humanistes ont disparu depuis longtemps du monde globalisé actuel mais c'est devenu un problème fonctionnel car les m2 sont de plus en plus chichement comptés et on travaille dans une réelle promiscuité, un bruit permanent, des odeurs, des interpellations permanentes , bref le souk. Il est impossible d'avoir un entretien téléphonique soutenu tant le dérangement à 1 mètre, voire 50 cm est permanent notammant à cause de super-managers hyper-kinétiques qui jaillissent de leurs cagibis individuels pour vous harceler de questions stupides . Tout cela fonctionne quand même grâce aux PC et aux email qui permettent de s'isoler dans une bulle virtuelle. J'ai aussi appris à ne plus regarder, ni écouter les gens pour me protéger, ils font partie d'un contexte visuel ou sonore au même titre que les plantes ou les armoires ou une muzak de supermarché. Du coup, on est aussi productif qu'une poule de batterie industrielle dont la production n'est pas de qualité moindre qu'une boule bio élevée en plein air selon ce que nous chantent Que choisir etc. On peut dire, de quoi se plaignent ses bureaucrates, c'est moins pire pour eux que dans le monde des usines où la promiscuité et le bruit on connait depuis des générations . Vrai mais ce qui est usant moralement et nerveusement , c'est le côté physiquement très statique . En fait, plus que des aspects financiers ou de sécurité, open space est mon motif Nř1 de dégoût du monde de l'entreprise et je préfère que mes enfants travaillent dans une ferme ou dans un atelier que dans ce monde orwellien.

342 & Exclu dans l'inclusion

343 & J'ai expérimenté les deux, et depuis janvier je suis seule dans mon petit bureau et je suis bien! Pas de coup de tél à longueur de journée, parler fort, chauffage à 30 et pas d'aération jamais, raconter sa vie privée au tel toujours, impossible de se concentrer! confidentialité préservée, et quant j'ai envie de parler je vais voir les autres collègues ou ils viennent me voir! La liberté vraiment !

344 & Ces effets étaient pourtant prévisibles, non ? Comment se fait-il que l'open space se soit malgré tout diffusé à ce point ?

345 & Étonnant, des humains qui pensent à court terme... Ou pas :x

346 & La "bonne réaction" du personnel assujetti à ces dispositif, c'est sans aucun doute d'évaluer le management sans ménagement, et donc de prendre *toutes* ses décisions et autres inventions pour ce qu'elles sont, le plus souvent dictées par des effets de pouvoir. Evidemment, il faut simuler l'intérêt pris au travail, et conserver une bonne dose de sens critique en face de ce qu'on ne peut pas changer dans l'immédiat. Après tout, le boulot, ce n'est pa

347 & ....] Evidemment, le boulot, ce n'est pas une religion, sauf si on est curé, pasteur, rabin ou imam. Alors, comment voulez vous que les "jeunes embauchés" (j'allais écrire "recrues") apprécient l'esprit maison, surtout si on leur fait comprendre ce que l'on attend d'eux sans véritable contrepartie autre que légale (salaire, conventions collectives, droit du travail..). Je ne suis plus en activité aujourd'hui, mais j'ai constaté que mettre en concurrence des gens qui ne peuvent pas -c'est une évidence- mener à bout des projets un peu complexes et supposant une bonne entente aussi bien verticale que horizontale, c'est aller à l'échec certain. Mettre dans la même arène ceux qui par fonction doivent d'abord coopérer, et de façon disons "amicale", c'est provoquer des oppositions, voire des inimitiés qui feront capoter toute espèce de réalisation où les rôles individuels (et les capacités aussi, souvent!) sont complémentaires et où la seule "concurrence" devrait se situer dans le domaine de la sociabilité non de la recherche du pouvoir. C'est pour cela qu'un "promu" à des fonctions managériales est souvent celui qui est inapte aux relations horizontales -ou obliques- serrées, le rôle du manager consistant (apparamment?) alors seulement à vérifier que les feuilles de temps sont remplies à la bonne date avec les imputations permises par la hiérarchie (cest un peu caricatural mais contient quand même un fond de vérité).

348 & Avant d'être en télétravail j'étais en Open Space. Pas idéal non plus, d'autant que pour m'aider à mémoriser lorsque j'ai beaucoup de choses qui m'arrivent en même temps, ou quand je dois appliquer une méthodo un peu compliquée, ou tout simplement quand je suis fatigué, qu'il y a du bruit et que je dois me concentrer, j'ai tendance à parler tout seul. Quand on fait ça en Open Space on surprend pas mal de regards méfiants et de sourires en coin autour de soi. Et là on se sent très seul, bien plus seul que chez soi sans personne autour.

349 & Ma grande boîte a créé un super espace ouvert, les uns sur les autres, bruyant, difficilement gérable... L'extension était prévue à d'autres sites... Marche arrière toute, démissions en chaîne et en plus de très bons... Congès maladie inexplicablement en hausse, jusqu'à une enquête par la DRH. Bilan on va réinstaller des mini- cloisons... La connerie humaine est limitée seulement par les budgets, ouf !

350 & Les "open space" font perdre en efficacité et en productivité. Quand a tenter d'obtenir d'un contact un document confidentiel par téléphone dans ces conditions, ce n'est même pas la peine d'y penser. L'open space signifie aucune possibilité de se concentrer dans de bonnes conditions, aucun moyen de vraiment travailler efficacement. Mais cela apprend à faire semblant, à perdre du temps pour se reposer sans en avoir l'air ... Le bruit, les appels tél des uns et des autres, c'est un stress continue qui épuise et réduit considérablement la productivité. Rien ne vaut les bureaux individuels quand on cherche la productivité et l'efficacité.

351 & Je vois bien l'application de ce machin dans les locaux de la dgse ou seulement de la dcri, voire de n'importe quelle entreprise privée hi-tech un peu sensible : bonjour les fuites ..., sauf si on veut les organiser.savamment et sciemment.

352 & Le DG SODEXO = ńil a fait preuve d'exemplarité, admire le directeur de projet. Il a un bureau à porte ouverte totalement transparent.ż Je travail là bas et c'est faux, Michel Landel DG de Sodexo a son bureau au dernier étage. Dans un coin bien à part. Ce n'est pas de l'open space. Il faut vérifier vos sources.

353 & ça dépend des open space , ça dépend de ce qui est fait. S'ils sont très petits et que beaucoup de gens téléphonent c'est pas top. Mais s'ils sont grands , bien fait , et que les gens passent pas leur temps à telephoner ça peut être bien. On voit les autres , on voit du monde , on voit passer les belles nanas ...Moi j'aime bien. Mais c'est vrai que souvent c'est bruyant , et ça c'est carrement contre-productif.

354 & les belles nanas, pour les mater, installez-vous à la terrasse d'un café. Votre réflexion est pathétique.

355 & Beaucoup d'image tellement précises dans ce texte. l'open space m'a moi aussi tué. @Tom : 8 personnes, je n'appelle pas cela un open-space. Chez moi, un openspace = 200 personnes

356 & Je me suis peut-etre mal exprime. Pour rentrer dans le detail, on est une cinquantaine a mon etage, avec 7 ilots de tables de 7-8 en "open space" et on partage une grande piece (avec, a cote, des salles de reunions a reserver selon les besoins -de confidentialite ou presentations de projets, par exemple). Et ce a chaque etage. Pas besoin d'etre des centaines pour parler "d'open space", c'est juste une methode d'agencement ou les bureaux ne sont pas separes dans des pieces differentes, c'est tout. Et cela peut etre bien fait, chaque equipe avec son ilot par exemple. 200 personnes a un meme etage agencés n'importe comment, bien sur que c'est la cohue, c'est bien pour ca que je parlais des open spaces "a taille humaine" dans mon titre :)

357 & Ils ont même prévu la rampe de lancement pour les suicidaires.

358 & Marre de ceux qui passent et se croient obliger de serrer la louche des 40 (si, si) présents dans l'open space, des facheux, des fachés, des maniaques, des gueulards au téléphones, des conversations vie privée, vie publique, des vannes nulles, d'être forcé de tout capter sans pouvoir se concentrer ou se décontracter, du con qu'on vous impose en remplacement du voisin supersympa, du forum permanent qui pourrit la vie.

359 & T'es tout tristounet. Explique moi ce que c'est un open space et le grand Dodo te répondra si tu as raison de tirer dans le tas..

360 & Ca a toujours existé, sauf qu'avant on n'appelait pas ça "bureaux open space". A présent on a la manie de fourrer de l'anglais partout pour rendre modernes des trucs très ringards. On y travaille mal, il y a trop de bruits et d'agitations. Pas pratique ni très poli non plus pour recevoir des visiteurs. On travaille mieux dans un bureau individuel. Je suis persuadé que le confort est un élément essentiel pour une meilleure performance. Enfin ça dépend du travail que l'on fait, mais parfois on a besoin de s'isoler pour se concentrer ou

simplement parce qu'on a une conversation confidentielle.

361 & ... à moins que ce soit en France. Open Space, prononcé avec l'accent français, ça pourrait en être hilarant si ce n'était pas pathétique.

362 & Dans quel monde évoluez vous pour ne pas parelr "Open Space". ... Sortez de votre bulle

363 & huhuhuhu encore un on parle comme ca en haut lieu dans nos boites excuse nous de parler de nos boites telles qu'elles sont Rassure toi, je sens bien ton préjugé arrivé, on parle aussi bien francais que toi en sortant de plus le net c'est mondial!!! alors tu trouveras surement d'autres mots d'autres langues dans l'avenir... bon allez afk cya

364 & Hé, y en a marre de vos anglicismes. "Open space" et "full open space", je ne comprends pas. C'est bien un truc de journaliste, on se la raconte un peu et ça fait classe d'utiliser l'anglais. Je suis bilingue et alors? Je m'efforce de toujours parler un français correct. Faites des efforts !

365 & Ce n'est pas un truc de journaliste c'est la réalité des entreprises aujourd'hui que de parler ce "franglais". Open space est dans le langage courant au même titre que "weekend" ou "planning".. Ce type de débat, intellectuellement surement intéressant, me parait totalement hors sujet .. Bon courage ...

366 & Le journaliste ne fait que transcrire un usage pour ainsi dire ancré dans les moeurs. Comme le montrent très bien Des Isnards et Zuber dans leur bouquin, dire "plateau" ou "espace ouvert" à la place d'"open space" ferait ringard, pas dans le coup, pas intégré. Nul, quoi. La novlangue des bureaux est un rite et un totem, il est très difficile de lutter contre. D'ailleurs, pour aller dans votre sens, plus les gens maîtrisent mal l'anglais, et plus ils affectionnent le franglais, dont le côté techno donne l'illusion d'être compétent et efficace. D'autant que sorti du bureau, leur français est digne d'une vache espagnole. Au fait, l'anglais des affaires, c'est comment par rapport à l'anglais de Milton ?

367 & Vous avez souvent entendu "je travaille dans un espace ouvert"??? les langues sont faites pour échanger et s'échangent allègrement des mots, des expressions. Devinez de quelle langue proviennent les mots suivants: cabas, bougie, jaquette, mesquin, nénuphar, nuque... allez, je vous mets sur la voie, comptez de 0 à 9

368 & à l'apparition de ces trucs,on les appelait "bureaux paysagés" c'etait pas mal. Mais les gens adorent, non pas l'anglais, mais le franglais. en l'espace de 20 ans notre mémé est devenue une mamie...etc

369 & moè mami coté maternel mémé coté paternelle au choix j'ai 41ans...

370 & moi j'ai fais les deux j'étais chez Orange en Open space, c'était génial car c'est une entreprise très cool, on avait assez d'espace pour avoir de l'intimité, je connaissais

presque tout le monde, c'est vrai qu'on deteste plus facilement les collegues qui font trop de bruits ou les tops models qui se dehanchent toute la journée . Mais c'était très plaisant Par contre je suis a la defense en ce moment dans un bureau de 2 personnes, c'est l'enfer, on s'epie toute la journée et si une n'est pas de bonne humeur, elle communique a l'autre cette mauvaise humeur, le pire c'est qu'on se dit pas bonjour puisqu'on ne se connait qu'a travers les mails.

371 & des grognasses qui se déhanchent serait un régal pour mon open space hourdé de testostérone

372 & autosurveillance, ipod, concours de mail, gueulard de service qui fait toutes ses affaires persos avec le tel du bureau et qui saoule tout le monde, devoir se concentrer avec plusieurs types qui parlent ou téléphonent à côté, problème de clim : toujours un frileux ou un irradié pour changer la clim sans demander rien à personne, le store qui fait travailler tout le monde dans l'ombre car celui à côté de la fenêtre ne vois rien sur son pc à cause des reflets (normal) chez nous les n 1, 2 3 ... ont des bureaux à l'ancienne...:s bref tout dans ce livre est ma réalité quotidienne

373 & aux compétitions de craché de noyaux d'olive. Bobo champion ile-de-France 2002-2005

374 & Apparemment, il y en a un paquet qui sont (ou ont des collegues qui sont) incapables de vivre en societe. Je bosse en petit open space (ilots de 6-8 personnes) et il n'y a absolument aucun probleme de voisinage, d'espionnage (sauf si vous souffrez de paranoia je suppose), etc. Le bureau est plutot calme et detendu. Pas besoin de se deplacer pour demander une info a un collegue, il est a cote/en face de vous! C'est meme tres sympa et ameliore clairement les relations boss/employe (et entre employes) puisque celui-ci n'est pas dans une tour d'ivoire mais a la meme enseigne que vous. Apres, les coups de telephones sont plutot rares dans mon departement et je concois que pour certaines activites style phoning cela puisse devenir rapidement trop bruyant. Mais bon rien n'empeche les managers d'utiliser leur cerveau et de voir si l'usage de l'open space est adapte a leur boite/departement ou non.

375 & J'ai connu l'open space notamment dans une petite boite familiale sauce start up (say yeah !). Dans la maison du patron, avec son chien, son frère, sa bagnole, sa femme, sa maîtresse (si si !), 3 ou 4 collaborateurs. Cette promiscuité infernale a fini par me taper sur les nerfs. Ma résistance ne tenait plus qu'à un fil. Difficile de décrire en quelques lignes le climat étriqué - qui s'auto-justifie comme fun, jeune et moderne. Ce que j'y perdais : la motivation, l'efficacité, le sérieux, la concentration (le sommeil aussi). Depuis lors, j'ai décidé de me mettre en indépendant, je travaille moins (mais mieux), je gagne beaucoup moins : assez pour couvrir mes besoins (sobres). Je continue à travailler entre autres avec eux - mais à distance : surtout de chez moi. Le travail que je faisais là-bas en une journée dans la ruche infernale, je peux le concentrer en moins d'une ¡ journée. Je vie selon mon rythme

376 & Moi jeune diplomée, j'ai toujours connu les stages plus ou moins en open space. La je viens de commencer un job de chargée de dvt dans une petite boite on est 6 dans le bureau et devinez quoi, je suis a la table la plus proche de ma responsable (qui est la chef d'entreprise) , c'est affreux, je deteste ca ! les premiers jours je n'osais meme pas prendre le telephone pour appeler a l'etranger var je sentais l'observation, alors que quand je suis seule, je suis bcp plus a l'aise meme si mopn accent et ma grammaire sont loin d'etre parfait. Et je vous dis pas les "Est ce que t'as fait ca, penser a ca tu peux m'aider sur ca...) CA m'enerve au plus au point. Une fois alors que j'ecrivais un long mail (c'est vrai a une amie) elle me balance avec un pti sourire "il est bien long ton mail !". Franchement je prefere encore avoir des collegues du meme niveau que moi dans le meme bureau, je vous assure qu'avoir son boss c'est juste affreux..D'ailleurs je saute de joie interieurement quand elle part en rendez-vous et l'ambiance est plus détendue (oui parce que j'ai oublié de dire que c'etait une colérique et une nerveuse..)

377 & Je conseillerais aux jeunes de lire "Surveiller et punir" de Michel Foucault. Ils comprendraient mieux la rationalité de l'open space, et ils comprendraient aussi que la mathématique financière, ce n'est pas tout.

378 & Open space encore une connerie à l'américaine.

379 & Soit l'open office conduit à un auto flicage du personnel soit çà conduit à tout savoir sur la vie privée des uns et des autres et çà ne favorise en rien le travail... En tant que chef d'une entreprise j'en ai rien à faire de la vie privée de mon personnel, il est là pour bosser et je le préfère dans l'intimité de leur bureau... s'ils ont besoin de se confier à d'autres çà ne me pose pas de problème ce qui me pose problème c'est la disparition de la vie privée...

380 & Et si le système n'était pas complètement mauvais (maxi 12/15 personnes) mais que nous n'étions tout simplement pas prêt car trop intéressé à adapter ou à calquer son comportement sur celui des autres : soit nous souhaitons ne pas en faire plus (faut pas exagérer en plus il gagne 20 de plus que moi le salaud) soit nous souhaitons sortir du lot (fayot) ! Tout est question d'ego et de respect de l'autre : car l'herbe est toujours plus verte chez son voisin ! Je remarque plus facilement que mon collègue arrive plus tard que moi qu'il termine à point d'heure (je suis pas là pour le voir) - qu'il n'a pas de travail lorsque je suis débordé que l'inverse - qu'il parle fort quand j'ai besoin de calme quand je suis au téléphone avec mon banquier (ça fait pas sérieux) :-))) En résumé tout le monde s'observe mais ne voit que sa réalité, celle qu'il veut bien voir (tu parles pas aujourd'hui tu fais la gueule - Nan je travail !). Tout le monde passe des coups de fils perso, surf sur internet pendant son travail mais c'est plus facile à accepter quand on ne le voit pas car dans ce cas on ne juge que le résultat final, l'efficacité de chacun et pas sa méthode et son organisation. Arrêtons de créer des mini-espaces complémentaires pour s'isoler car à force d'être tous dans des placards il va bientôt falloir retourner travailler sur le plateau déserté pour être enfin au calme et ne pas avoir à partager cet espace impersonnel et exigu. Cloisonnons intelligent, créons des open-space visuel et non sonore, cloisonnés par des vitres et sans portes afin que chacun puisse s'inviter pour partager, travailler - pour le

reste il y a la machine à café et le resto. Rassemblons les équipes exerçant le même métier afin qu'il existe une véritable synergie par groupe de 6/8 maxi. Si tous les patrons, cadres ou supérieurs hiérarchiques, les chefs les vrais, étaient dans une même pièce beaucoup de problèmes seraient plus rapidement réglé ou leurs expériences respectives seraient un force pour l'entreprise. Arrêtons de mélanger cadre, assistante, responsable, secrétaires, commerciaux... Nous ne sommes pas tous identiques, nous subissons tous un stress différent, nous ne mangeons pas tous les mêmes rillettes... Une seule chose est sûr nous avons tous besoin d'un espace de travail efficace, motivant, épanouissant pour sortir de cette p..... de crise ! YesUcan you too :-)

381 & ça fait 5 ans que je bosse dans un open space à Berlin, et c'est infernal: - le bruit constant est extenuant - les batailles quotidiennes l'été pour savoir qui a ouvert la fenetre avec la clim à donf - les batailles quotidiennes l'hiver pour fermer les fenetres parce que moi, j'ai froid après 5 mn quand il fait -2řC dehors mais qu'apparement, il y en a qui trouve qu'il fait encore trop chaud dedans - je ne vous parle pas de la guerre des stores (au moindre rayon de soleil, qq'un baisse le store parce que c'est trop lumineux) - Au niveau des lampes on approche de guerre de sécession (j'allume, qq'un éteint 2 mn après) - il y a toujours un troupeau de grumeaux ou une grognasses pour décider de discuter avec leurs camarades qui sont assis à 6 metres, donc ils hurlent, et ce bien sûr pile à l'instant de ma conférence téléphonique avec mes collègues aux USA ou de ton entretien téléphonique avec ton manager - il y a toujours quelqu'un pour regarder sur ton écran - même quand tu bosses, c'est désagréable au possible (je trouve ça effrayant: j'ai développé une sorte de 6eme sens pour ça; je dois être un mutant) - impossible de peter tranquille donc j'angoisse les jours ou il y a du choux à la cantine - etc... moralité: l'enfer c'est les autres (pas con, ce Sartre)

382 & tu n'es pas un mutant, tu n'es pas seul...

383 & Le problème principal de l'open space est que se sont plutôt les mauvais élément qui tirent l'open space vers le bas, plutôt que les bons éléments qui le tirent vers le haut. Quand un mec bosse et qu'il voit ses collègues se la couler douce, il aura tendance à lever le pied. Combien d'heures de travail gachées par ces open space !!! Mais c'est mieux pour la drague, on peut s'envoyer des petites oeillades pleines de sous entendus ...

384 & Merci d'éviter les barbarismes anglais dans vos textes en français. Ou alors écrivez tout en anglais!!

385 & chut

386 & pour l'avoir vécu pendant 7 ans à La Défense, dans deux tours différentes (Winterthür et EDF), l'open space, c'est le petit enfer au quotidien. Du genre à vous dégoûter de votre entreprise et de son mode de management.

387 & je préfère un bureau perso. On peut y péter comme on veut...

388 & bon argument pour le bureau fermé, mais le taux de méthane y sera plus concentré... L'avantage de l'open space, c'est que ça se dilue plus vite. arfff !!

389 & L'open space,c'est comme à l'école sans aucune intimité et une surveillance hiérarchique. On infantilise les gens, on ne les fait pas progresser au final. Ou est le bénéfice ? L'idéal sont les bureaux de 2 ou 3 personnes qui me semblent un bon compromis. Les secrétaires regroupées entre elles dans un bureau par exemple, ce qui favorise les échanges mais laissent un peu d'intimlité et évite le phénomène Big Brother qui fait que les gens n'ont pas envie de venir travailler le matin. L'Open Space, pas franchement dynamisant et efficace au final. D'ailleurs aucune étude n'a montré que les sociétés en Open Space avaient de meilleurs résultats que des sociétés à disposition classique.

390 & pour ces open spaces: des economies pour le patron, une surveillance accrue sur les salariés mais des conditions nulles pour un travail efficace : ce bruit permanent empêche une reflexion logique, car il y a toujours des grandes gueules qui ne peuvent parler doucement et crient dans le téléphone, font des plaisanteries grasses en croyant que choquer c'est faire de l'humour open space ? ou comment apprendre à détester la vie collective.. et les managers fous qui copient servilement le monde anglo saxon, sans meme prendre le temps de voir si c'est compatible avec les traditions européennes... naze de chez naze !

391 & Pour ma part, ayant fait les 2, retourner dans un petit bureau, à 1 2 ou 3, j'aurais du mal, je trouve pas çà terrible, ou alors il faut tomber sur 1 ou 2 bonnes personnes. Je m'endors dans un petit bureau ! Au moins un open space, çà vit ! On voit plein de monde, on discute plus facilement, on voit les jolies filles passées.. ;-) Si j'ai un coup de fil important, on a des petites salles ou on peut s'isoler. De plus, si flicage il y a, il ne se fait pas/plus visuellement mais sur votre pc. Le flicage visuel, c'est un peu de la parano, les gens/managers ont autre chose à faire... Mais ce n'est que mon avis, on est d'accord.

392 & Ces gens-là n'ont jamais été pions dans une permanence de collège. Sinon il sauraient que la productivité dans le brouhaha est à peu près égale à zéro. Une certaine culture du pauvre fait défaut à nos managers.

393 & je hais les open spaces. Ils me rendent associale. J'en arrive à détester mes collègues. Mon rêve : travailler chez moi pour être tranquille.

394 & La petite chose que les concepteurs de ces open-spaces ont dû oublier, c'est la fait que la communication est un exercice qui demande un reflexion sur soi. Soit elle demande des petits aménagements où se retrouver avec soi-même pour faire le pendant à l'ouverture vers les autres. Soit un mode de fonctionnement "communicatif" est trouvé et devient une sorte d'automatisme de comportement de 8h à 22h...qui souvent se fissure de manière assez dramatique en cas de stress.Dans ce genre de bureau les "managers" communicants ont souvent des pétages de plombs memorables...ces p...d'actifs survitaminés.Ca ne marche pas ce genre d'espace. Etendu à la vie sociale en général, celà revient à dire que le foyer ne sert à rein et que pour le bien de l'évolution de l'humanité, il faut loger tout le

monde dans des lofts 24h/24. “l'effet cocktail ” est tout l'inverse de ce que présenté dans l'article: Cet “effet” illustre la capacité du cerveau huamain à comprendre une discussion au milieu d'un brouhaha grâce à la combinaison des bribes d'info auditives et visuelles (bouche principalement) de votre interlocuteur. Effectivement, cet effet est déficient si vous ne voyez pas votre interlocuteur (ex. telephone). L'effet cocktail ne se résume pas au fait d'élever la voix dans un brouhaha pour se faire comprendre. Merci de prendre en compte cette précisioncar ce brave effet cocktail est souvent utilisé...à toute les sauces. (oui, je sais). Bonne journée à tous dans vos open-spaces!

395 & Entre le grand plateau de plusieurs dizaines de personnes au moins, et le “bureau d'équipe” où on logera de l'ordre de cinq personnes au plus, il y a quand même de fortes différences. Dans le second cas, les inconvénients par rapport au “bureau fermé à deux” sont assez faibles... le bureau individuel pour tous (pas seulement pour les chefs et les “managers”) étant quand même quelque chose d'assez rare de toute façon, même dans les entreprises qui ont gardé des bureaux cloisonnés.

396 & Pour un décideur, cette organisation est dans le vent (panurge n'est pas mort), de plus le directeur financier en remet une couche (m2 en moins, baisse des charges). Dans ma très grande boîte, au bout de 8 mois, c'est très positif...., pour la badgeuse qui voit de plus en plus de salariés, même cadres se sauver de ce bordel dès que l'heure arrive. Le temps que le diplodocus comprenne on aura bien cassé la machine...

397 & c'est un vrai cirque dans le quel tout le monde hurle dans son téléphone pour entendre son auditeur. La cruche qui vient d'obtenir son CDI se permet de se meler de ce qui ne la regarde pas. elle prépare déjà son projet de carrière... De plus dans celui-ci (hé oui, c'est le 2ème que je connais) il n'y a pas de vitre pour se protéger du face à face. Super, on n'entend pas les clients (malgré le casque) donc on hurle et se postillonne joyeusement ses microbes (surtout en hiver) sur la figure. c'est complètement nul, mais c'est dans l'aire du temps n'est-pas ? et la mode est un éternel recommencement.

398 & Bah l'article est assez clair : c'est un rêve de manager : flicage mutuel entre grouillots, autocensure et impression de productivité. Que ça soit hypar méga chiant pour les grouillots, bah... S'y sont pas contents ils peuvent aller voir ailleurs, hein...

Résumé

Le développement du Web 2.0 a donné lieu à la production d'une grande quantité de discussions en ligne. La fouille et l'extraction de données de qualité de ces discussions en ligne sont importantes dans de nombreux domaines (industrie, marketing) et particulièrement pour toutes les applications de commerce électronique. Les discussions de ce type contiennent des opinions et des croyances de personnes et cela explique l'intérêt de développer des outils d'analyse efficaces pour ces discussions.

L'objectif de cette thèse est de définir un modèle qui représente les discussions en ligne et facilite leur analyse. Ce modèle a été implémenté en partie pour satisfaire aux besoins du projet "Conversession" développé pour une société start-up soutenue par CREALYS, incubateur d'entreprise de la région Rhône-Alpes. L'objectif de la société était d'extraire des connaissances et d'analyser les discussions lors de débats sur internet.

Dans cette thèse nous proposons un modèle basé sur des graphes. Les sommets du graphe représentent les objets de type message. Chaque objet de type message contient des informations comme son contenu, son auteur, l'orientation de l'opinion qui y été exprimée et la date où il a été posté. Les liens parmi les objets message montrent une relation de type "répondre à". En d'autres termes, ils montrent quels objets répondent à quoi, conséquence directe de la structure de la discussion en ligne.

Avec ce nouveau modèle, nous proposons un certain nombre de mesures qui guident la fouille au sein de la discussion et permettent d'extraire des informations pertinentes. Les mesures sont définies par la structure de la discussion et la façon dont les objets messages sont liés entre eux. Il existe des mesures centrées sur l'analyse de l'opinion qui traitent de l'évolution de l'opinion au sein de la discussion. Nous définissons également des mesures centrées sur le temps, qui exploitent la dimension temporelle du modèle, alors que les mesures centrées sur le sujet peuvent être utilisées pour mesurer la

présence de sujets dans une discussion. La présence de l'utilisateur dans des discussions en ligne peut être exploitée soit par les techniques des réseaux sociaux, soit à travers notre nouveau modèle qui inclut la connaissance des auteurs de chaque objet message.

La représentation d'une discussion en ligne de la manière proposée permet à un utilisateur de "zoomer" dans une discussion. Une liste de messages clés est recommandée à l'utilisateur pour permettre une participation plus efficace au sein de la discussion.

De plus, un système prototype a été implémenté pour permettre à l'utilisateur de fouiller les discussions en ligne en sélectionnant un sous ensemble d'objets de type message et naviguer à travers ceux-ci de manière efficace. Les techniques actuelles de fouilles de textes ou d'opinions ont été intégrées dans ce prototype, ce qui montre comment le modèle proposé rend possible la fouille d'une discussion en ligne.

Mots-clés : discussions en ligne, opinion mining, fouille de données d'opinion, text mining, fouille de texte, réseaux sociaux, systèmes de recommandation, modélisation, forum.