The Revieval of Subject Analysis: A Knowledge-based Approach facilitating Semantic Search

Sebastian Furth[1], Volker Belli[1], and Joachim Baumeister[1,2]

[1]denkbares GmbH, Friedrich-Bergius-Ring 15, 97076 Würzburg, Germany
[2]University of Würzburg, Institute of Computer Science,
Am Hubland, 97074 Würzburg, Germany
{sebastian.furth,volker.belli,joachim.baumeister}@denkbares.com

Abstract. Semantic Search emerged as the new system paradigm in enterprise information systems. However, usually only small amounts of textual enterprise data is semantically prepared for such systems. The manual semantification of these resources typically is a time-consuming process. The automatic semantification requires deep knowledge in Natural Language Processing. Therefore, in this paper we present a novel approach that makes the underlying Subject Indexing task rather a Knowledge Engineering than a Natural Language Processing task. The approach is based on a simple but powerful and intuitive probabilistic model that allows for the easy integration of expert knowledge.

Keywords: Subject Indexing, Document Classification, Semantic Search

1 Introduction

Historically, Subject Analysis and *Subject Indexing* [10,1] had been a rather manual task, where librarians or catalogers tried to index large corpora of documents according to a given set of controlled subjects. A more technical but prominent example for large scale Subject Indexing is the web catalog from the early Yahoo times, where websites had been indexed with certain topics. Regardless of which medium was used, catalogers typically tried to determine the overall content of a work in order to identify key terms/concepts that summarize the primary subject(s) of the work. An indexing step enabled in-depth access to parts of the work (chapters, articles, etc.). Therefore, the item was conceptually analyzed (what is it about?) and subsequently tagged and cataloged with subjects from a controlled vocabulary.

Nowadays, *Semantic Search* [8] applications belong to the state of the art in Information Retrieval. In contrast to traditional search engines ontologies are used to connect multi-modal content with semantic concepts, which can then be exploited during the retrieval to improve search results. Therefore, users of Semantic Search applications typically formulate their search queries as semantic concepts. Then a retrieval algorithm might expand the query considering ontological information. Finally, a look-up method maps the concepts to actual search results using an index from concepts to information resources.

With the growing amount of information manually maintaining catalogs or indices became almost impossible. However, catalogs and indices are typically built for a specific problem domain. For many domains formal knowledge in form of ontologies exists and comprises decent amounts of terminology and relational information. Thus, we describe a novel approach for automatic Subject Analysis that allows for the easy integration of formal domain knowledge. We have built an intuitive probabilistic model that makes Subject Analysis not a Natural Language Processing but rather a Knowledge Engineering task. Therefore, the approach allows that domain experts can control the analysis by expressing their knowledge about relations between concept classes and the importance of certain document structures.

The remainder of the paper is structured as follows: Section 2 formally defines the Subject Analysis problem and discusses related work. In Section 3 we present our Knowledge-based Subject Analysis approach. Section 4 describes experiences made with our approach in industrial scenarios. We conclude with a discussion of our approach in Section 5.

2 Problem Description

2.1 Controlled Vocabularies

The fundamental requirement for Subject Analysis and the subsequent Subject Indexing is the existence of a controlled vocabulary. Historically, a controlled vocabulary defined the way how concepts were expressed, provided access to preferred terms and contained a term's relationships to broader, narrower and related terms. Nowadays, such information is typically modeled by standardized ontologies [9,13,12], where terms are embedded in complex networks of concepts covering broad fields of the underlying problem domain. Typical examples are ontologies powering semantic enterprise information systems. In such systems users interact using concepts that are company-wide known and valid. An increasing amount of companies maintain corresponding ontologies as they are the key element for the interconnection of enterprise systems and data [18]. If such ontologies do not exist, the construction is usually very reasonable under cost-benefit considerations, as they support not only semantic information systems but are also a vehicle for the introduction of more elaborate services like Semantic Autocompletions [11] or Semantic Assistants. In this paper, we formally define a controlled vocabulary as follows:

Definition 1 (Controlled Vocabulary). *A controlled vocabulary is an ontology $O = (T, C, P)$ that contains a set of terms T that are connected to a set of concepts C. Concepts $c \in C$ are connected to other concepts using properties $p \in P$.*

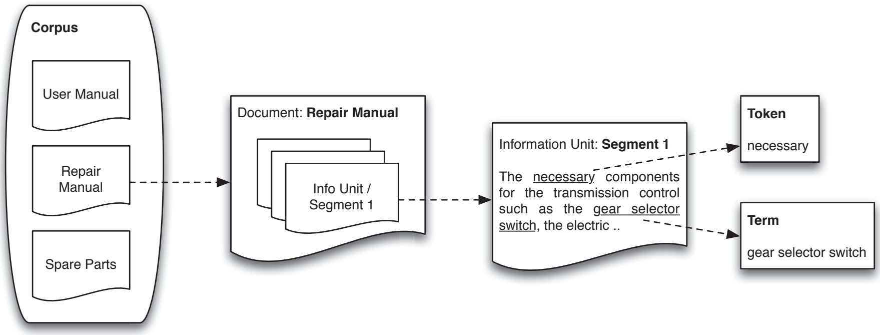
Corpus
User Manual
Repair Manual
Spare Parts
Document: **Repair Manual**
Info Unit / Segment 1
Information Unit: **Segment 1**
The necessary components for the transmission control such as the gear selector switch, the electric ..
Token
necessary
Term
gear selector switch

Fig. 1. Partitioning a corpus to information units.

2.2 Subject Indexing

Assuming that such an ontology/controlled vocabulary exists the task is to examine the subject-rich portions of the item being cataloged to identify key words and concepts. Therefore existing textual resources must be partitioned to sets of reasonable *Information Units* [6,3,17,4] (see Figure 1). Then the task can be defined as follows:

Definition 2 (Subject Indexing from information units). *For each Information Unit $i \in I$ find a set of concepts $C_i \subseteq C$ from an ontology O that describe the topic of the corresponding text best.*

An information unit i has an associated bag of term matches M_i, i.e. a list of terms from a domain ontology/controlled vocabulary that occur in a particular information unit. Given the bag of term matches the task can be specialized as follows:

Definition 3 (Subject Indexing from bags of words). *Given a bag of term matches M_i determine the underlying topics in the form of a set of concepts $C_i \subseteq C$ from an Ontology O.*

The availability of formalized domain knowledge is usually a valuable support factor for tasks that cover certain aspects of a problem domain [14,15,16]. We claim that this is also true for Subject Indexing where the selection of topics can profit from formalized background knowledge. Thus, the integration of domain knowledge in the annotation mechanism becomes a critical success factor and the task can be further refined as follows:

Definition 4 (Subject Indexing with background knowledge). *Given a bag of term matches M_i determine the underlying topic in the form of a set of concepts $C_i \subseteq C$ from an Ontology O considering the domain knowledge contained in Ontology O.*

2.3 Related Work

Topic Analysis is a relatively wide field of research and is strongly influenced by Document Classification and Document Clustering approaches. Notable approaches exist in particular among latent methods, i.e. topics are not expressed in form of explicit concepts but as a set of key terms. Prominent examples are Latent Dirichlet Allocation (LDA)[2] and Latent Semantic Analysis (LSA)[5]. Regarding the deduction of explicit topics Explicit Semantic Analysis [7] is a well-known approach.

3 Probabilistic Subject Analysis

In the following, we will first present a basic probabilistic model that is based mainly on weighted semantic relations between terms and concepts. The model can be tailored to integrate expert knowledge for a certain domain specific controlled vocabulary. The basic model will be extended in order to also consider document characteristics, like important document structures (e.g. headlines) or formatting information (e.g. bold text).

3.1 Basic Probabilistic Model

The basic probabilistic model is founded on observable text/term matches, relations between these terms and potential topics (concepts) and a strong independence assumption between all features. The model connects the features as follows:

1. Starting from a text match $match \in M_i$ in an information unit $i \in I$ the model derives potentially corresponding terms $t \in T$.
2. The model *optionally* weights the term $t \in T$ with respect to the covering document structure of the corresponding text match $match \in M_i$, e.g. term occurrences in headlines might be more important.
3. Given a term $t \in T$ the model looks for concepts $c \in C$ that can be described with this term, i.e. which concepts have this term as label and how specific is this label.
4. The concepts $c \in C$ derived from the model on basis of the text/term match $m \in M_i$ might have relations to topic concepts $topic \in C_i$ with $C_i \subseteq C$. The model exploits ontological information for the derivation of topic concepts $topic \in C_i$ from observed (term) concepts $c \in C$ resulting in a topic probability for a text/term match.
5. The derived topic probabilities for each text match $match \in M_i$ get aggregated in order to compute the overall topic probabilities for an information unit $i \in I$.

Given a bag of term matches M_i for an information unit $i \in I$, we realized steps (1) to (3) by computing the topic probabilities for each text/term match $match \in M_i$:

$$\mathbf{P}(\mathbf{topic} \mid \mathbf{match}) = \alpha * P(topic \mid c) * P(c \mid t) * P(t \mid match). \tag{1}$$

Therefore, we consider the confidence of a term match $P(t \mid match)$, i.e. the probability of a certain term t given a textual match *match*. Additionally we take the specificity $P(c \mid t)$ of a term t for a certain concept c into account. Unique labels have the maximum specificity of 1.0. The relevance of the concept in focus c for a topic concept *topic* is $P(topic \mid c)$. This relevance gets computed on basis of ontological information between both concepts. The relevance is maximum if both considered concepts are equal (identity). Finally, we use the constant prior α to express the linguistic uncertainty that a certain topic is not meant given a certain term match. This avoids that one perfect term match pretends other topics to get more important, i.e. it regulates how many related term matches are necessary in order to outperform one perfectly matching term. We then compute the topic probabilities for an information unit $i \in I$ (step 4) on basis of the topic probabilities for each term match $match \in M_i$:

$$\mathbf{P}(\mathbf{topic}) = 1 - \prod_{match}^{M_i} (1 - P(topic \mid match)). \tag{2}$$

The result is a set of topics $topic \in C_i$ with associated probabilities that express how well a certain topic fits to the terminology observed in the information unit. This computation assumes independence between the term matches in M_i according to Bayes' Theorem. The independence assumption might not perfectly reflect reality but is a sufficient approximation in this application scenario.

3.2 Extended Probabilistic Model

The basic probabilistic model can be extended, such that it also considers distinctive document characteristics as valuable background knowledge. In many specialized publications like technical documents or textbooks document structures indicate the underlying topic or support at least the discrimination of multiple topic candidates. Typical examples are headlines or formatted text (italics, bold, underlined).

The basic probabilistic model uses a constant prior α that expresses the linguistic uncertainty that a topic is not meant by a certain term match. We extend the basic model, such that the prior is not constant but depends on the document structure where the term match was observed. Therefore, document structures get weighted according to their importance for the deduction of a topic for an information unit. Assuming that for each document structure a weight w exists (default 1.0) the value for the prior α is computed as follows:

$$\alpha_{\mathbf{adaptive}} = 1 - (1 - \alpha_{constant})^{w}. \tag{3}$$

This procedure also allows to discriminate document structures that are inappropriate for the topic deduction, e.g. references/links to other documents.

3.3 Knowledge Representations and Derivation of Probabilities

The preceding sections introduced a simple but powerful and intuitive probabilistic model for Subject Analysis. However, the primary target remains that the Subject

Analysis of large document corpora becomes rather a Knowledge Engineering than a Natural Language Processing task. Therefore, the proposed probabilistic model allows for the easy adaptation to characteristics of a domain specific controlled vocabulary and the corresponding corpus of documents that shall be subject indexed. The following section describe the knowledge-based adaptation, i.e. the definition of basic conditions for the derivation of probabilities.

Term Confidence $P(t \mid match)$ The term confidence $P(t \mid match)$ expresses how certain a text match is actually a term occurrence. The computed confidence depends on the quality of the text match. A perfect match, i.e. the text match *match* is equal to the term t results in the maximum confidence of 1.0. The usage of fuzzy string matching techniques like order independent matching, stemming etc. might lower the confidence of term matches. Therefore, implementations of the presented probabilistic approach should allow for the configuration of different fuzziness levels and adjust the confidence accordingly.

Term Specificity $P(c \mid t)$ Given a term the model must derive all concepts $c \in O$ that can be described by this term. The model must also express how specific a term is for a concept $P(c \mid t)$, i.e. handle ambiguous terms like “apple” which can be the name of a company or a fruit. In the context of technical documents, we might encounter terms like “nut”, “engine” or “screw” that are very ambiguous and thus unspecific. Therefore, the specificity of a term must be distributed over all potential concepts. In the simplest case the specificity can be distributed equally over all concepts. Unambiguous terms always have a specificity of 1.0. However, experts’ knowledge might be used to prefer certain concepts. This might be useful if some concepts of an ontology are not applicable, e.g. because components they represent are not included in certain machines.

Concept Relevance $P(topic \mid c)$ Then, given a concept the model must be able to determine how relevant it is for certain topics $P(topic \mid c)$. The procedure is always the same and is explained by the example of technical documents. In technical documents the occurrence/observation of a concept describing a component might be relevant for a couple of concept topics: (1) machine functions relying on this component, (2) parent components or (3) the component itself.

In general, we assume that the relevance of a concept for a topic decreases the larger the distance between both concepts is in the underlying ontology. However, experts’ might know that in certain situations (documents) the occurrence of a concept is much more indicative for specific topics than for others. For example in operator manuals component terms might also indicate functions while they typically do not in repair manuals because usually an operator wants to "operate a function", whereas a technician usually wants to "repair a component".

For the calculation of the concept relevance distances between concepts and topic concepts are extracted/queried from the ontology. Expert knowledge can be used to weight these distances according to the properties $p \in P$ involved. This way background knowledge regarding the relevance of certain concepts under certain circumstances can be included in the model. Finally the weighted distances between the concept in focus c and the topic concept *topic* get transformed to a probability. We propose the usage of a normalized sigmoid function to avoid overestimation of the distance. The parameters β and γ can be used to control the sigmoid function and thus the overall importance of the concept relevance:

$$\mathbf{P}(\mathbf{topic} \mid \mathbf{c}) = \frac{1 + e^{(-\beta) * \gamma}}{1 + e^{(distance - \beta) * \gamma}} \tag{4}$$

Linguistic Uncertainty $\boldsymbol{\alpha}$ In the basic probabilistic model the parameter α is constant. In the extended model the parameter α can be adjusted, such that it can prefer or discriminate term occurrences in certain document structures. Therefore, domain experts can define weights w for certain document structures (default 1.0). Values for w greater than 1.0 prefer, values smaller than 1.0 discriminate terms in certain structures respectively. During the computation of the value for the adaptive linguistic uncertainty $\alpha_{adaptive}$ an implementation has to consider the value accordingly.

4 Extended Example

An exhaustive and thorough evaluation of the presented approach is subject to future work. However, we have already applied the probabilistic model in an ongoing industrial semantification project with promising results. In the following we briefly describe the key aspects of the case study.

4.1 The data set

In the case study the task is to semantify a given corpus of technical documents provided in PDF format. The corpus comprises several thousand pages of technical information, spreaded over different documents like operator manuals, functional descriptions or repair and maintenance instructions. The semantification partitions the PDF files to reasonable segments (information units). Then, each information unit is subject indexed with respect to an existing ontology.

The ontology contains information about the hierarchical structure of components in the corresponding machine as well as functional connections between components (see Figure 2 for a simplified visualization). Labels are attached to all concepts.

4.2 Parametrization of the Probabilistic Model

The probabilistic model has been parametrized to incorporate existing domain knowledge. Therefore, we used the tailoring possibilities described in Section 3.3 as follows:

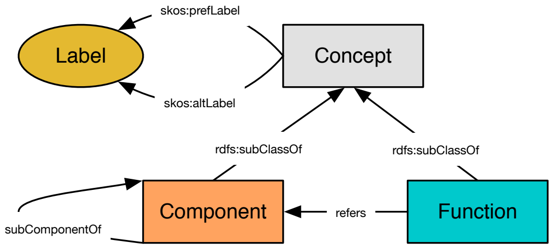

Fig. 2. Simplified visualization of the domain ontology.

- **Term Confidence** $P(t \mid match)$**:** We allowed order independent lookups without decreasing term confidences. Matches that had only been possible due to stemming have been discriminated.
- **Term Specificity** $P(c \mid t)$**:** We have distributed the specificity equally over all concepts, i.e. if a term is attached to two concepts, the specificity of the term is 0.5 for both concepts.
- **Concept Relevance** $P(topic \mid c)$**:** For operating manuals we slightly preferred the `refer` property, in descriptive manuals the `subComponentOf` property respectively.
- **Linguistic Uncertainty** $\alpha_{adaptive}$**:** We defined weights w greater than 1.0 for headlines and captions, i.e. prefered term matches occurring in the heading of sections and the descriptions of images.

A formal evaluation has not yet been performed. However, experts reviewed the derived topics and confirmed a noticable improvement over a previous implementation based on Explicit Semantic Analysis.

5 Conclusion

In this paper we presented a novel approach for automatic Subject Indexing, i.e. the indexing of information units with respect to a controlled vocabulary/ontology. The presented approach is based on a simple but powerful and intuitive probabilistic model. We claim that this approach does not require training data but facilitates the easy incorporation of experts' domain knowledge and thus is highly adaptive. The adaptiveness through experts' knowledge makes automatic Subject Indexing rather a Knowledge Engineering than a Natural Language Processing task. Thus, large scale semantification of enterprise corpora becomes possible. The approach has not yet been evaluated thoroughly. However, its application in industrial case studies yielded promising results.

Besides an exhaustive evaluation future directions include the addition of learning methods. Therefore, we consider incorporating latent approaches as preprocessors to adjust concept relevances based on term frequencies in the underlying corpus. Additionally, we plan to investigate whether simulated annealing can be used to learn weights w for document structures. We also plan to consider background knowledge about (hierarchical) connections between document structures in the model, e.g. the consideration of neighbour or parent segments' topics.

Acknowledgments

The work described in this paper is supported by the Bundesministerium für Wirtschaft und Energie (BMWi) under the grant ZIM ZF4172701 "APOSTL - Accessible Performant Ontology Supported Text Learning".

References

1. Albrechtsen, H.: Subject analysis and indexing: from automated indexing to domain analysis. Indexer 18, 219–219 (1993)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Borkar, V.R., Deshmukh, K., Sarawagi, S.: Automatic segmentation of text into structured records. In: Mehrotra, S., Sellis, T.K. (eds.) SIGMOD Conference. pp. 175–186. ACM (2001), http://dblp.uni-trier.de/db/conf/sigmod/sigmod2001.html#BorkarDS01
4. Choi, F.Y.Y.: Advances in domain independent linear text segmentation. In: ANLP. pp. 26–33 (2000), http://dblp.uni-trier.de/db/conf/anlp/anlp2000.html#Choi00
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391–407 (1990)
6. Furth, S., Baumeister, J.: Semantification of Large Corpora of Technical Documentation. IGI Global (2016), http://www.igi-global.com/book/enterprise-big-data-engineering-analytics/145468
7. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on artificial intelligence. vol. 6, p. 12 (2007)
8. Guha, R., McCool, R., Miller, E.: Semantic search. In: Proceedings of the 12th international conference on World Wide Web. pp. 700–709. ACM (2003)
9. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. (eds.): OWL 2 Web Ontology Language: Primer. W3C Recommendation (27 October 2009), available at http://www.w3.org/TR/owl2-primer/
10. Hutchins, W.J.: The concept of 'aboutness' in subject indexing. In: Aslib Proceedings. vol. 30, pp. 172–181. MCB UP Ltd (1978)
11. Hyvönen, E., Mäkelä, E.: Semantic autocompletion. In: The Semantic Web–ASWC 2006, pp. 739–751. Springer (2006)
12. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax (Feb 2004), http://www.w3.org/TR/rdf-concepts/
13. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. (2009), http://www.w3.org/TR/2009/REC-skos-reference-20090818/
14. Milne, R., Nicol, C., Trave-Massuyès, L., Quevedo, J.: TIGER: Knowledge based gas turbine condition monitoring. AI Communications 9(3), 92–108 (1996)

15. Padma, T., Balasubramanie, P.: Knowledge based decision support system to assist work-related risk analysis in musculoskeletal disorder. Knowledge-Based Systems 22(1), 72–78 (2009)
16. Puppe, F., Buscher, G., Atzmueller, M., Huettig, M., Buscher, H.P.: Clinical experiences with a knowledge-based system in sonography (sonoconsult). In: Workshop on Current Aspects of Knowledge Management in Medicine (KMM05), Proceedings 3rd Conference Professional Knowledge Management - Experiences and Visions, Kaiserslautern, Germany (2005)
17. Reynar, J.C.: Statistical models for topic segmentation. In: Dale, R., Church, K.W. (eds.) ACL. Association of Computer Linguistics (1999), http://dblp.uni-trier.de/db/conf/acl/acl1999.html#Reynar99
18. Stephens, S.: The enterprise semantic web. In: The Semantic Web, pp. 17–37. Springer (2007)