HNews: an enhanced multilingual hyperlinking news platform

Diego De Cao, Daniele Previtali, and Roberto Basili

University of Roma Tor Vergata, Roma, Italy
{decao,previtali,basili}@info.uniroma2.it

Abstract. In this paper, we describe the *HNews* platform, a Web-based system addressing the general problem of aggregating and enriching news from different sources and languages. In the indexing stage, the news items gathered from RSS feeds or video streams are analyzed through Information Extraction tools. Their topical category information and the Named Entities mentions are recognized and used to create semantic metadata so to enrich the information available for each news item. Moreover, a robust unsupervised Word Sense Disambiguation algorithm is applied to the available texts that are thus further semantically annotated. This is used to align news items in different languages, such as Italian and English, and support cross-lingual search. As a result, advanced search features, such as cross-lingual or typed entity-based queries, are enabled in HNews. In this paper, we also present the browser, making use of a spatial metaphor for the arrangement of the retrieved news. It enables to capture different aspects such as the "semantic" similarity among news, or the timeliness of individual news items as well as their relevance with respect to an incoming user query.

1 Introduction

As globalization emerges, information access across language boundaries is becoming a critical issue. The World Wide Web has become accessible to more and more countries and technological advances overcome the network, interface and computer system differences which are constraints to information access. In particular the World Wide Web is becoming a major "media" for news delivery (e.g. broadcasting) and content creation. Consequently, it has now an increasing appeal for users the ability to search news from different sources, media and in different languages. The application of Information Retrieval techniques to the problems raised by news aggregation in such heterogeneous scenarios is becoming a crucial technological challenge. Currently, the major technology enabling the information access across different sources is represented by the *News Aggregator* software. Aggregators reduce the time and effort needed to regularly check websites for updates as well as for creating a unique integrated information space or a "personal newspaper". Correspondingly, every aggregator has explored the way of integrate some Information Retrieval capability in order to reduce the effort to satisfy real user information needs.

Research on ad hoc retrieval systems focus on a variety of methods, these ranging between strongly lexicalized, statistical methods for relevance modeling to more semantic-oriented techniques, usually based on deeper levels of text analysis and language processing paradigms (such as parsing or semantic disambiguation processes). In the former, so-called shallower, approaches, documents are retrieved through simple matching mechanisms between texts and queries; ranking according to relevance is the side effect of statistical models of terms co-occurrences in texts. Semantic approaches attempt to exploit at a certain degree the syntactic and semantic information made available by linguistic analysis. In the attempt of reproducing some levels of text understanding, a meaning surrogate of the input text is obtained including also semantic (sometime syntactic) indexes. These are metadata that restrict the potential interpretations of the texts and are supposed to improve the retrieval accuracy. For example, a smaller number of false positives are expected, as constraints at the semantic level can be imposed to re-rank candidates documents. The technology that supports the detection, disambiguation and formalization of meaningful information, nowadays called semantic metadata, from unstructured texts is *Information Extraction* (hereafter IE), as it has been studied since early 90's ([1]).

Notice that the availability of semantic information is useful in particular in cross-linguistic scenarios, whereas strongly lexicalized statistical methods are not of much help: they in fact cannot be used to retrieve documents that are expressed in a language different from the one used to query the IR system. In these cases, beyond merely accepting extended character sets and performing language identification, the information retrieval systems should also be able to provide help in locating information across the language boundaries. Moreover access to distributed information is complex also due to the heterogeneity of the sources and to diversity of interests, expectations and purposes of the target retrieval processes. Heterogeneity here characterizes:

- Data typologies, as sources of information are characterized by different media and content types.
- Data formats, as even the same content can be made available through a media according to different levels of granularity and quality. Formats may highly vary across and within archives.
- Contents, as the source information is not characterized by a specific knowledge domain but is spread across heterogeneous semantic dimensions.
- Languages, as even an individual structured archive may well include documents expressed in different natural languages.

The major media channel for news is represented by television. In previous work, the problem of extracting of semantic metadata from broadcasted TV and radio news has been discussed and a corresponding system, RitroveRAI, is presented [2]. It makes use of human language technologies for IE over multimedia data (i.e. speech recognition and grammatical analysis of incoming news). The *HNews* system, presented in this paper, represents an extension of RitroveRAI, as it integrates the indexing of video news with the gathering and annotation of news from different Web sources. News derived from newspaper portal in the

Web are characterized by texts that are less noisy than speech transcriptions. As a consequence, in HNews, a set of different natural language processing modules is applied and a comprehensive enrichment of individual news through semantic metadata is obtained. The next section presents the overall structure of the applied indexing process, while Section 3 describes the search interface. The section 4 concludes this work by discussing applications and extension of the system.

2 Enriching Web news through semantic metadata

RSS is a family of web feed formats used to publish frequently updated news contents in a standardized format and is adopted by of the web newspapers in publishing timely updates. The *HNews* platform exploits these news aggregator standards, and collect news updates from independent RSS feed sources. However, as rich metadata are required to improve the quality of the retrieval process, the limitations of current RSS protocols must be carefully handled in a system such as HNews. The idea of applying IE to contents requires that the RSS-supported data gathering process is followed by an in-depth analysis according to a family of advanced natural language processing tools. Unfortunately, the RSS feed of most newspapers makes available only a summary of individual news, on which a too small scale linguistic analysis is possible. The summary in fact is usually very short and insufficient to perform accurate extraction (e.g. sense disambiguation is more complex when shorter contexts are targeted). As an extension of a classical news aggregator, *HNews* provide crawling capabilities, use RSS links to access the corresponding complete news contents, i.e. full Web articles pages. A specific RSS processor for individual newspaper sources has been developed at this purpose.

Once the full textual content is made available, a cascade of NLP tools is applied to extend contents with semantically rich metadata. The most interesting among these metadata is the set of entities, such as the places, locations, organizations and temporal expressions mentioned in the news item. Accordingly, a statistical Named Entity Recognition module is applied to detect and compile lists of NEs (i.e. persons, cities, companies) that will be part of the metadata related to the targeted item. The applied NER tool is further described in Section 2.1). A relevant information used to organize and retrieve news items is the editorial class, i.e. the topical category corresponding to the news content. While these are usually made available by the different providers through the RSS format itself, the reference classification scheme adopted by the various sources is highly varying and differs from one another. In order to determine a unified and comparable scheme, news from different sources are classified by the HNews system into a set of predefined editorial categories, inspired by previous work on this problem [2]. The supervised statistical classification process adopted in HNews is described in section 2.2. Further analysis is carried out to disambiguate sense of relevant words through a Word Sense Disambiguation (WSD) stage, as discussed in section 2.3. WSD here is applied especially to natural language queries in order to support cross-linguistic search. Finally, the comprehensive

set of information about an underlying news item (i.e. title, text content, named entities, topical category and words senses) are indexed through a well-known engine (i.e. Lucene [3]). The above process is applied to the textual components of Web pages, although according to a separate independent workflow (as discussed in the description of the RitroveRAI system [2]), HNews is also able to index TV broadcasted news, whenever the segmented news and their speech transcriptions are made available (see for example, [4]) according to the workflow described in [2]. The rich form of semantic metadata in HNews allows to integrate semantically typed information in order to make heterogenous sources and different languages coexist and support flexible forms of querying and conceptual aggregation. While Section 3 will discuss the navigation capabilities and the resulting information space made available by the HNews dedicated GUI, the rest of this section will describe the main HLT processing stages applied by HNews.

2.1 Statistical Named Entity Recognition

The Named Entity Recognition (NER) task aims at the identification of all named locations, persons, organizations as well as dates, times, monetary amounts or other numerical expressions that appear in free forms in a text. NER is a crucial step for the enrichment proposed by HNews as it highly improves the performance of news aggregation. A reference system for NER is certainly IdentiFinder [5], that makes use of a variant of a Hidden Markov model to identify names, dates or numerical quantities. In the original proposal, states of the HMM are designed to correspond to the above classes. There is a conditional state for "not a token class". Each individual word is assumed to be either part of a specific pre-determined class or not part of any class. According to the definition of the task, one of the class labels or the label that represent "none of the classes" is assigned to every word. IdentiFinder uses word features, which are language-dependent, such as capitalization, numeric symbols and special characters, because they give good evidence for identifying tokens. A version of an HMM-based NER has been designed at our Lab, and it has been trained against annotated Web material in Italian for the major categories of *people*, *locations*, *organisations* and *dates*. The resulting NER module is applied in the HNews workflow both for news and query processing. Moreover, given the extremely noisy nature of speech transcriptions, two different HMM-based recognizers are adopted against the texts and the segments of TV broadcasted. While performances close to 87% accuracy are obtained over standard textual input, a not negligible performance drop is observed over transcribed speech material, where 70% is the reachable accuracy on average.

2.2 News Categorization

Text categorization has been traditionally modeled as a supervised machine learning task [6]. In *HNews*, a simple yet efficient model, i.e. Rocchio, is applied. The model, described in [7] is a profile based classifier, where a specific

cross validation process allows to optimize at individual class level and obtain performance close to state of the art systems (e.g. Support Vector Machines). Given the set of training document R_i , classified under the topics C_i (positive examples), the set $\overline{R_i}$ of the documents not belonging to C_i (negative examples) and given a document d_h and a feature f, the Rocchio model [8, 7] defines the weight Ω_f of f in the profile of C_i as:

$$\Omega_f^i = max \left\{ 0, \frac{\beta}{|R_i|} \sum_{d_h \in R_i} \omega_f^h - \frac{\gamma}{|\overline{R_i}|} \sum_{d_h \in \overline{R_i}} \omega_f^h \right\} \tag{1}$$

where ω_{fh} is the weight of the feature f in the document d_h. In equation 1, the parameters β and γ control the relative impact of positive and negative examples and determine the weight of f in the *i-th* profile. In [8], values β=16, γ=4 have been first proposed for the categorization of low quality images. These parameters indeed greatly depend on the training corpus and different settings of their values produce significant performance variations.

Notice that, in equation 1, features with negative difference between positive and negative relevance are set to 0. This represents an elegant feature selection method: the 0-valued features are irrelevant in the similarity estimation. As a result, the remaining features are optimally used, i.e. only for classes for which they are selective. In this way, the minimal set of truly irrelevant features (those having a weight of 0 for all the classes) can be better captured and removed.

In [7], a modified Rocchio model is presented that makes use of a single parameter γ_i as follows:

$$\Omega_f^i = max \left\{ 0, \frac{1}{|R_i|} \sum_{d_h \in R_i} \omega_f^h - \frac{\gamma_i}{|\overline{R_i}|} \sum_{d_h \in \overline{R_i}} \omega_f^h \right\} \tag{2}$$

Moreover, a practical method for estimating the suitable values of the γ_i vector has been introduced. Each category in fact has its own set of relevant and irrelevant features and equation 2 depends on γ_i, for each class i. Now, if we assume the optimal values of these parameters can be obtained by estimating their impact on the classification performance, nothing prevents us from deriving this estimation independently for each class i. This results in a vector of γ_i each one optimizing the performance of the classifier over the *i-th* class. The estimation of the γ_i is carried out by a typical cross-validation process. Two data sets are used: the training set (about 70% of the annotated data) and a validation set (about 30% of the remaining data). First the categorizer is trained on the training set, where feature weights (ω_f^d) are estimated. Then profile vectors Ω_f^i for the different classes are built by setting the parameters γ_i to those values optimizing accuracy on the validation set. The resulting categorizer is then tested on a separate test set. Results on the Reuters benchmark are about 85%, close to state-of-art more complex classification models ([7, 9]). Extensive discussion on the performances reached over different benchmarks is reported in ([7]).

Score	Instance ID	Category	Title	Date
0,202	Ans_1345651810	Editoria, Stampa e Mass Media	Agcom, diffida a Tg1 e richiamo a Tg4 e Studio Aperto	22-ott-2010
0,185	CdS_1405379650	Politica, Partiti, Istituzioni e Sindacati	Antigua, D'Alema:«Premier si dimetta» E Romani attacca Report: puntata odiosa	19-ott-2010
0,179	CdS_1666402036	Politica, Partiti, Istituzioni e Sindacati	Il Cavaliere: a questo punto si ritiri il Lodo	23-ott-2010

Fig. 1. Categorization of Web news in *HNews*

A typical example of the obtained results is reported in Figure 1, that shows the results of a retrieval session in HNews: the third column reports the topical classification of each retrieved news item. While the last two entries are derived from the "Corriere della Sera" (CdS) portal, and their topic label is already available, i.e. "Categoria: Politica", the first originates from Ansa, and it is missing of any topic label. Column 3 in the Figure reports the automatically labels assigned by the HNews classifier, that in the last two cases confirms the original CdS classification. Notice how while the first two news are dealing with similar topics, their focus is different and this is very well reflected by the classifier.

2.3 Applying Word Sense Disambiguation for Query expansion in CLIR

Lexical ambiguity is a fundamental aspect of natural language. Word Sense Disambiguation (WSD) investigates methods to automatically determine the intended sense of a word in a given context, according to a predefined set of sense definitions. These are usually provided by a reference semantic lexicon. The impact of WSD on IR tasks is still an open issue and large scale assessment is needed. Unsupervised systems are certainly very interesting as for their applicability to non English, i.e. resource poor, languages. While state-of-art systems are usually supervised, their porting to other languages is mostly expensive as large scale resources are needed. For this reason, unsupervised approaches to inductive WSD are very appealing. In the framework of the HNews architecture, we adopt a network based model of WSD based on WordNet, as discussed in [10]. In [11], a variant of the the PageRank algorithm, called personalized PageRank is presented. WordNet is assumed as a network of senses and a random walk model of its links is defined. Then, sentences, or entire texts, are used to initialize the state of the WordNet network, and the stable state of its "random walk" is assumed as the posterior statistics across the senses of the targeted words. While the approach can be applied either to individual words or to entire sentences, in [10] it has been shown that a distributional approach can improve the personalized PageRank disambiguation algorithm, both in accuracy and time complexity. The initial state is obtained as determined by a topical expansion of individual sentences through the use of Latent Semantic Analysis [12]: sentence is firstly mapped into a vector in the LSA space and the closest words to the vector are retained and added to the sentence terms. The initialization of the network with this expanded lexicon provides the resulting PageRank to converge

first and to more accurate sense distributions. Details of the technique are discussed in [10]. In [10] a detailed evaluation of the adopted system for English is reported, moreover in [13] we have reported evaluation of the applicability of the WSD system to the Italian language. Results over the Senseval '07 benchmark are about 71,5% of F1 for the English. Instead, the Evalita '07 benchmark is used to evaluate the Italian language reaching about 52% of F1. Major lack is due to the difference of the WordNet version used between English and Italian language. For its application into HNews, a large collection of Web news, considered as the specific domain corpus, has been used to derive the LSA model where distributional evidence is represented. We first built the classical vector space model and then Latent Semantic Analysis is applied. Notice that the topical similarity across news are able to better characterize typical contents for the suitable word senses of all terms in a news item. When a sentence, a document or a query is input, first it is expanded with the set of its closest words in the LSA space. The expanded query is then used to trigger the personalized PageRank that provides the final preferences for senses of individual terms, e.g. the nouns, verbs or adjectives used in a query.

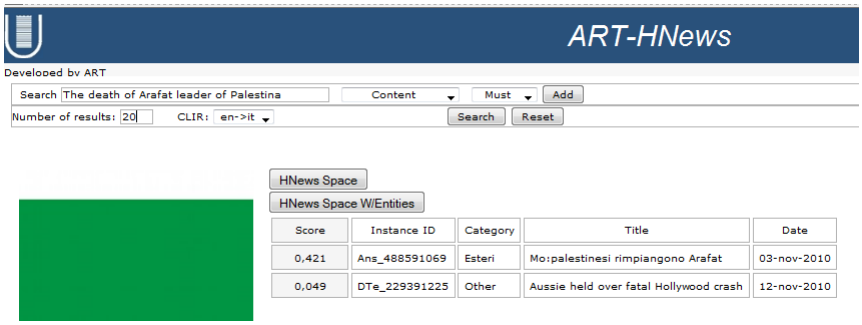

Fig. 2. CLIR in HNews: Italian and English news in responde to a query in English

While senses can be part of the document index, the very interesting aspect that characterizes the adoption of WSD in HNews is that it is an enabling factor for Cross Language Information Retrieval (CLIR). In fact, although WordNet [14] was developed for the English language, several versions for other languages, aligned with the English sense hierarchy, have been developed. MultiWordNet [15] is a WordNet for Italian that is strictly aligned with Princeton WordNet 1.6, at the synset level. A large number of English synsets are in fact put in correspondence with an Italian synset: words in this latter are thus synonyms each other while being specific "translations" of the words in the source English synset. The application of our WSD model to CLIR is thus straightforward. First, an English query is processed and its terms and Named Entities are extracted. Then nouns and verbs are disambiguated through our personalized PageRank and their WordNet senses are detected. Finally a translated query in Italian is obtained. It includes the original (language neutral) Named Entities as well as the Italian words obtained from the MultiWordNet synsets corresponding

to the selected senses of English words. In this way two versions (English and Italian) of the same query are obtained and documents written in both languages can be returned. In Figure 2, the response to the English query "*The death of Arafat, leader of Palestina*" is shown whereas the first hits include both news in Italian and English. Notice that scores are able to separate well meaningful from irrelevant news.

3 The retrieval interface

In a general IR scenario, the user interface must be able to support the user to submit queries to the system and trigger the navigation or browsing of the returned documents. An example of the browser interface is shown in Figure 3.

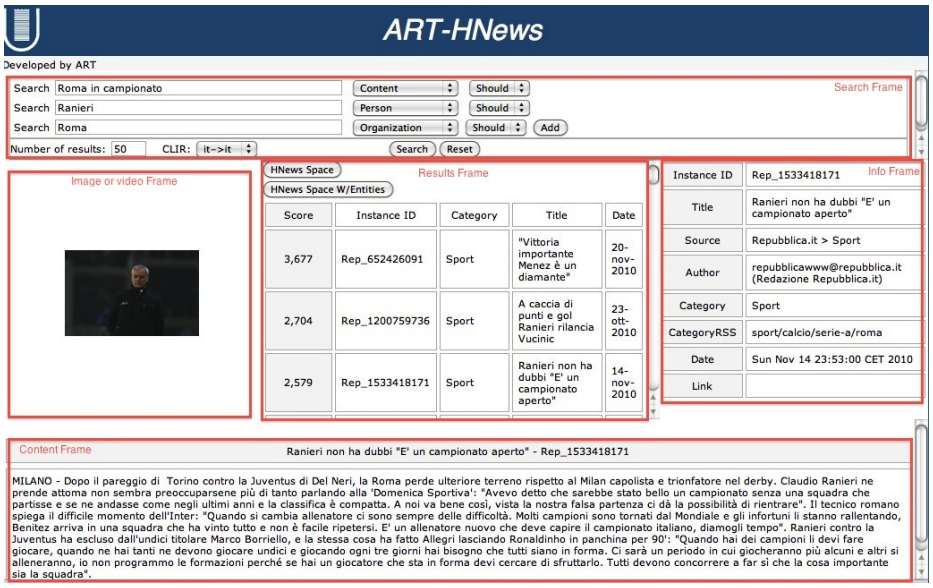

Fig. 3. *HNews* search interface

The interface is composed by five different frames, i.e. the top one, three in the middle and the footer one (see the red boxes in Figure 3). The top frame provides the query interface, were user can edit its queries as

- simple expressions, i.e. bag of keywords
- short texts in two languages that are analyzed by HLT tools
- boolean combination of simple queries

A variable set of constraints, i.e. individual simple queries, can be designed by the user, according to the different type of metadata considered. Relevant

fields of the indexes range in fact from the Topical Category, to full text content, or Named Entities. The query shown in Fig. 3 is the expression

```
(Content: "Roma in Campionato" AND Person:"Ranieri" AND
                    Organisation:"Roma")
```

that is "*Find all news that discuss the Roma team in the football league and Ranieri*, i.e. the current coach. It is also possible to specify whether one individual condition *must* or *can* be satisfied adding some further flexibility in the boolean combination of individual constraints (Fig. 3).

The middle frame includes three individual frames. In the central one the returned results for a query are shown, as already seen in Fig. 1. On the user click any result triggers multiple visualization actions. The middle left frame is used to show the video or the photos related to the selected news item. The middle right frame shows the metadata related to individual news items, such as publication dates and times, Web source or editorial category. The bottom frame, at the footer, changes according to individual selections in the returned news and it shows their textual contents.

Whenever an interesting news item has been found, the user can browse the Web, as links to the originating pages (at the source RSS portal) are made available. Moreover, the HNews platform also allows to support a spatial metaphor where an information space is shown to the user, centered in the selected news item in the set of the returned answers. The space is obtained through the mentioned Latent Semantic Analysis of the underlying collection. It aims at capturing the semantic relatedness between news (either Web or multimedia) as well as between news and Named Entities. A graph local to one selected news item can be in fact obtained by retrieving all news closer to it in the LSA space. The graph also expresses arcs among these news whenever close (i.e. similar) enough news pairs are involved. An example of the spatial view is shown in Figure 4. In the navigation tool, different visual layers are made available to capture different useful information:

- Every link represents the *similarity of a news pair* in the LSA space[1], so that the closer two news the more relevant they are for the same queries.
- Different font sizes are used to discriminate *timeliness*: the more recent a news item is, the larger its font will be. The resulting zooming effect tries to compensate *coverage* with *timeliness*: very old news will not limit visualization of more recent material, although being retrieved and shown
- Colors from green to red are used to represent *relevance* of a news item for an input query: green expresses more relevant, i.e. better, responses, while red is used for the worse ones.
- Shapes discriminate semantic types. While news (i.e. textual objects) are shown as colored boxes, ellipses are used to represent entities, and their semantic types, such as person vs. organizations. Notice that typed entities

[1] The Latent Semantic Analysis is the same performed for the WSD step and described in section 2.3

are represented in the LSA space as much as documents, so that their positioning in the network and their similarities are seemingly computed and depicted.

The resulting graphs have the desirable side effect of expressing a global view on the information space. Links express distances in the LSA space and this implies that news naturally organize into visual clusters, usually made of topically related materials. In the example in Figure 4, news aggregates form two major regions: the bottom one concerning mostly stories related to the "*economics*" class, while the upper right one concerns mainly "*sport*" topics.

As presented in section 2.3, the system is also able to search news in different languages as side effects of word sense disambiguation. In Figure 2, it is shown how a query can be submitted in English through a specific parameter (`CLIR: en -> it`) of the query frame (i.e. the upper frame). We already discussed the returned results that also include news from the Italian channels such as *Corriere della Sera* or *Ansa*. Notice that also the viceversa (i.e. query in Italian and documents in English) is currently supported.

4 Conclusions

In this paper ,the *HNews* system for semantic annotation and indexing of news from Web newspapers or TV broadcasts has been presented. A complex family of language processing tools is adopted for Information Extraction, i.e. the automatic recognition of different semantic information. Different models and approaches are exploited in HNews to extract rich metadata, such as Named Entities, topical categories or word senses. The resulting platform is also supporting cross-lingual queries and advanced boolean combinations of simple queries acting over texts and metadata. In particular, two languages are currently supported: Italian and English. News are downloaded from the Web on a continuous basis. Indexing proceeds from RSS feeds, so that published materials are available in almost real time. One of the mostly innovative aspects of the *HNews* system with respect to previous experiences in this area (e.g., the Prestospace system, [2]) is the browsing modalities offered, that integrates the search and semantic navigation functionalities. A quantitative model of semantic similarity is in fact defined over the rich set of metadata and connected graphs of news items and Named Entities in the information space are correspondingly obtained. These are quite effective to quickly focus on the information of highest relevance/interest, as their conceptual, rather than mere textual, nature is made explicit in the graph. HNews is the basis for future developments targeted to the support to the creation of a community of readers but also producers of news and other contents. In this way, the HNews portal would be directly reusable to support the gathering of user-generated contents. The exploitation of these latter for developing models of large scale, realistic opinion mining processes is the focus of future research enabled by the HNews system.

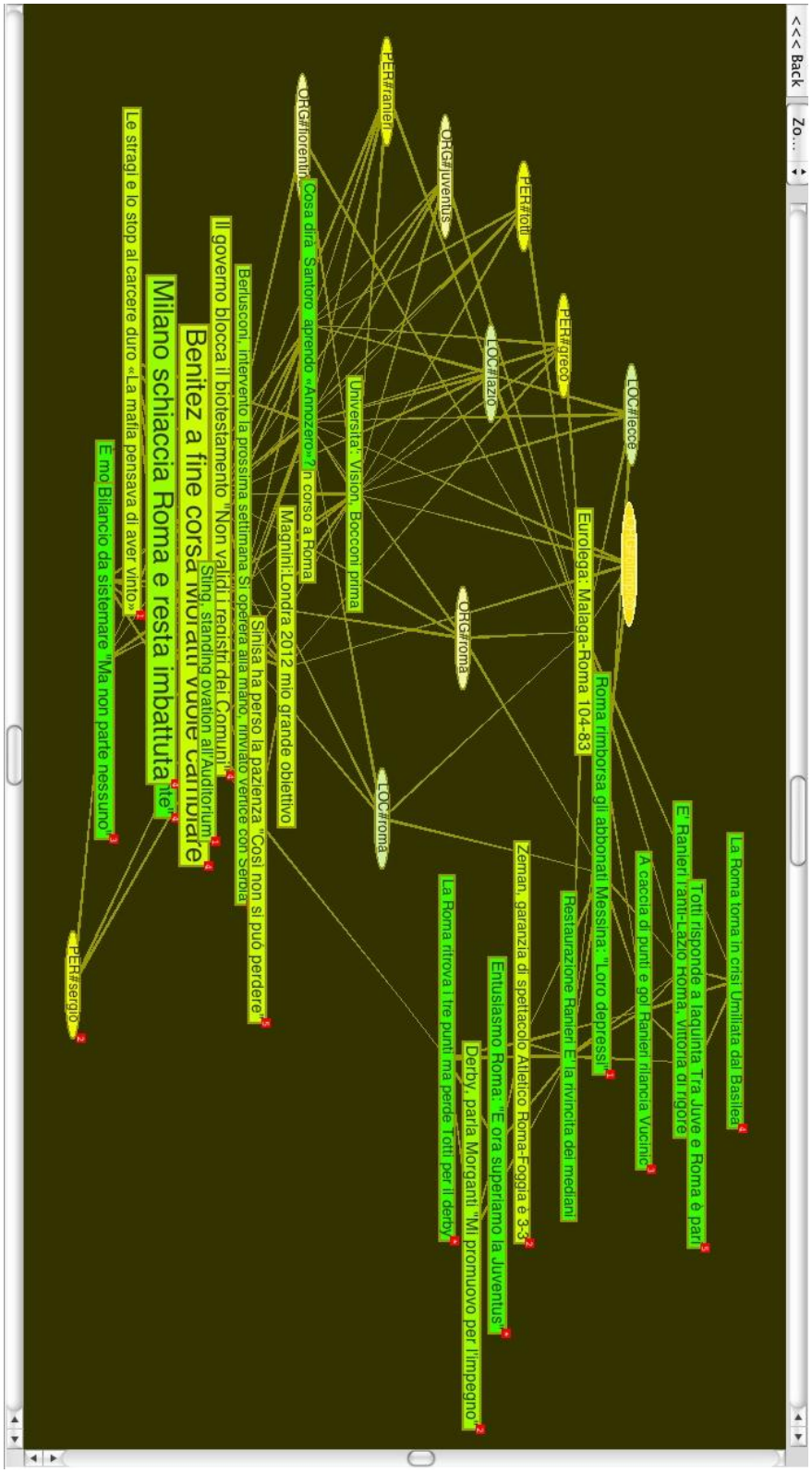

Fig. 4. *HNews* news navigator

References

1. Pazienza, M.T., ed.: Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, SCIE-97, Frascati, Italy, 14-18, 1997. In Pazienza, M.T., ed.: SCIE. Volume 1299 of Lecture Notes in Computer Science., Springer (1997)
2. Basili, R., Cammisa, M., Donati, E.: Ritroverai: A web application for semantic indexing and hyperlinking of multimedia news. [9] 97–111
3. Gospodnetić, O.: Advanced Text Indexing with Lucene. O'Reilly Media (2003)
4. Messina, A., Boch, L., Dimino, G., Bailer, W., Schallauer, P., Allasia, W., Groppo, M., Vigilante, M., Basil, R.: Creating rich metadata in the tv broadcast archives environment: The prestospace project". In: Proceedings of the AXMEDIS Conference. (2006)
5. Bikel, D., Schwartz, R., Weischedel, R.: An algorithm that learns what's in a name. Machine Learning Journal (1999)
6. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34**(1) (2002) 1–47
7. Basili, R., Moschitti, A., Pazienza, M.: Nlp-driven ir: Evaluating performance over a text classification task. In: Proceeding of the 10th "International Joint Conference of Artificial Intelligence" (IJCAI 2001), Seattle, Washington, USA (2001)
8. Ittner, D.J., Lewis, D.D., Ahn, D.D.: Text categorization of low quality images. In: Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval. (1995)
9. Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: The Semantic Web - ISWC 2005, 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, 2005, Proceedings. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: International Semantic Web Conference. Volume 3729 of Lecture Notes in Computer Science., Springer (2005)
10. De Cao, D., Basili, R., Luciani, M., Mesiano, F., Rossi, R.: Robust and efficient page rank for word sense disambiguation. In: Proceeding of TextGraphs-5: Graph-based Methods for Natural Language Processing, Uppsala, Sweden (2010)
11. Agirre, E., Soroa, A.: Personalizing pagerank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. EACL '09, Morristown, NJ, USA, Association for Computational Linguistics (2009) 33–41
12. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review **104** (1997) 211–240
13. De Cao, D., Basili, R., Luciani, M., Mesiano, F., Rossi, R.: Enriched page rank for multilingual word sense disambiguation. In: Proceeding of 2nd Italian Information Retrieval 2011 Workshop. (2011)
14. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller., K.: An on-line lexical database. International Journal of Lexicography **13**(4) (1990) 235–312
15. Pianta, E., Bentivogli, L., Girardi, C.: Multiwordnet: developing an aligned multilingual database". In: Proceedings of the First International Conference on Global WordNet, Mysore, India (January 21-25 2002)