Hypatia: an Expert System Proposal for Documentation Departments

Angel L. Garrido, Alvaro Peiro, Sergio Ilarri
IIS Department
University of Zaragoza
Zaragoza, Spain
{garrido, peiro, silarri}@unizar.es

***Abstract*—Nowadays, the vast amount of text-based information stored in organizations requires different approaches and new tools in order to manage it adequately. This paper presents Hypatia, a support expert system for documentation departments and regular users that exploits not only local information, but also external resources from the Web (e.g., Linked Data). The expert system uses different modules: Natural Language Processing (NLP) analysis, categorization, semantic disambiguation, Automatic Query Expansion (AQE), semantic search, summarization, knowledge extraction, and aggregation. Users can interact with the expert system in different ways, varying from giving very specific orders to writing a simple list of keywords. The latter method will require a previous interpretation before deciding the response of the system. The obtained results will benefit from semantic links referencing complementary data to improve both the information presentation and the data navigation.**

Keywords*—*Databases, Ontologies, Expert Systems, Semantic Web, Linked Data, Documentation Departments.

I. Introduction

In almost any public or private organization that manages a considerable amount of information, activities related to documentation can be found. To do this job, large companies have documentation departments, that have the responsibility of storing and managing the company's documents efficiently and effectively. However, the large amount of information in text format that organizations accumulate, both from internal and external sources, cannot be properly processed and documented by these departments (as a paradigmatic example, consider the large amount of documents and articles managed by the documentation department of a newspaper). The main reason is that most organizations assign few human resources to this task and the employees who carry out all these documentation activities have limited time to fulfil their assignments. The obtained results are subject to errors due to these conditions and to the subjectivity of the individuals.

Documentation is also an expensive and time-consuming task, whose activities can sometimes be automated, as human contribution is usually not relevant in these subtasks. In addition, some activities like tagging or classification can be very subjective, needing in some cases an explanation of the reasons motivating the actions applied. That is why a smart tool that automates documentation would help to improve the quality of the activities performed, as the applied reasoning strategies will be able to explain the outcome of the input.

The purpose of this paper is to propose an expert system called Hypatia, that helps documentation employees and regular users, using internal resources inside the company where it has been installed and external resources available in the Web. Hypatia offers different functionalities related to different tasks in the document analysis field: NLP analysis, categorization, semantic disambiguation, semantic search, AQE, summarization, knowledge extraction, and aggregation. The users can input keywords into the system, along with some optional features, in order to perform a more precise command. In case of a simple list of keywords, Hypatia has to disambiguate and decide which possibilities will offer to the user before performing the task chosen. Once it finishes, the expert system takes advantage of the semantic links which reference complementary data, to improve the navigation between the data and the obtained results. These results differ in their complexity, varying from simple Web links to visual representations and custom reports of the extracted information.

This paper provides three main contributions:

- Firstly, an expert system devoted to documentation departments has been developed. This system integrates state-of-the-art technologies oriented to information retrieval, extraction, and synthesis.
- Secondly, the system has been improved incorporating a methodology for obtaining information from the Web by exploiting Linked Data [1].
- Finally, the expert system has been evaluated with real users, in order to check the interest and the novelty of this proposal, with very good results.

This paper is structured as follows. Section II provides an overview of the related work. Section III explains the general architecture of the proposed solution and its inner modules. Section IV discusses the results of the evaluation of the proposal. Finally, Section V provides the conclusions and some lines of future work.

II. Related Work

Expert systems are highly-specialized software systems that can handle real-world problems, reaching conclusions similar to those that a human expert could obtain. However, documentation is a broad and challenging activity to tackle, since it has many open problems that are still being researched. Up to the authors' knowledge, there is no similar work regarding the definition of expert systems for the documentation field. Content and document management systems, such as Drupal, Joomla, Alfresco, or LifeRay, provide a partial commercial

solution to some tasks relevant for documentation departments, but they do not cover the whole set of activities that are attempted to achieve in this work. Since there is no similar work to compare with, this section is focused on different major tasks in the documentation field to describe and contextualize the state of the art of this paper.

Classification. Including documents in categories to facilitate searching and management is a key task in documentation departments. Automatic categorization can be performed by using software based on statistics and the frequency of words, and also by using machine learning systems [2]. However, these techniques need to be complemented by tools which deal with aspects related to NLP [3] in order to enhance their results. The aforementioned procedures feed from the know-how of the organizations, which can be made explicit and exploited thanks to the use of ontologies [4] and reasoners [5]. The output of the techniques can be related to the thesaurus [6] of each organization, so the final results can be expressed in terms of the vocabulary of each company. This tagging task can also be used on Web pages to improve the search and retrieval tasks, but it is still an open problem [7] due, among others, to the existing semantic and linguistic difficulties.

Information retrieval. Another key issue is to have a search system that facilitates access to the documents that are relevant to a given query. These data are usually stored in relational repositories that are queried by search systems that rely on keyword-based interfaces, and which do not consider the syntax and semantics of the data. In the context of keyword search on relational databases [8], most approaches retrieve data that exactly match the user keywords, but they do not consider the semantic contents of the keywords and their relationships. This can lead to losing information or to return needless data that are not interesting. This problem has received a great deal of attention and several approaches based on AQE solutions have been proposed to solve it. AQE is currently considered a promising technique to improve the information retrieval effectiveness and it is being adopted in commercial applications, especially for desktop and intranet searches.

Knowledge extraction. Relevant information of a certain topic from numerous different sources can be extracted by creating an automatic summary of all the data obtained. This summary is created by giving a score to each sentence of the text and selecting the ones with the highest scores. Classical methods use statistical criteria to detect sentences that contain more high-frequency terms [9], [10], while other recent works use positional criteria, such as the optimum position policy (OPP) [11]. Other coefficients, such as the Dice coefficient [12] and the Jaccard index [13], can compute the similarity between sentences to reduce the redundancy in the summary. On the other hand, metrics such as the Maximal Marginal Relevance (MMR) [14] also consider information novelty to score the sentences of the future summary. Other proposals work with multiple documents, using different approaches. For example, [15] uses a cluster centroid combined with other techniques, while [16] uses domain-independent techniques.

Aggregation. The creation of automatic reports is an activity that feeds from the other previously explained tasks: disambiguation increases the possibility to obtain the correct sense of the searched topic, and the extracted knowledge is the main input of the aggregation process. The extracted information can come from different sources that can be classified into semi-structured (or structured) sources and unstructured sources. The organization that provides the semi-structured sources usually offer an easy way to obtain most of the desired information. However, unstructured sources require analyzing natural language to extract the data. Recent works use event recognition and extraction [17] to aggregate the obtained events according to their similarity level.

III. Proposed Architecture

This section explains the general architecture of the proposed system. Hypatia contains three different working units plus an interface the users can interact with. Unless otherwise specified, the modules of these units have been and developed by the authors of this paper. Figure 1 shows these modules, along with the resources that they use. In the following, the next subsections describe the different existing units.

A. Language Unit

This inner unit provides three modules related to language analysis and manipulation. The remaining units of Hypatia use these modules to implement their internal functions.

NLP module. This is a library that encapsulates all the structures required to analyze plain input text. This module can work with different languages, as its architecture can support different NLP tool suites which can provide analysis for diverse languages. For English and Spanish, it uses Freeling[1], adapting its structures and functionalities to their counterparts in Hypatia. It implements four types of analysis, depending on the required morphological and syntactic information. The first one provides only the morphological analysis, giving for each word its lemma, grammatical number and gender, part of speech, etc. The rest of the analysis types offer an increasing amount of syntactic information, from a simple shallow parsing to a more complex rule-based dependency parsing, in which the module can distinguish between different grammatical relations in each sentence.

This module is widely used within the other components in Hypatia, including other libraries inside the Language Unit. It offers its functionalities through its own analyzers, as they include high-level functions to provide an easy access to the text analysis. The module has been tested in other environments and projects, both in Spanish and English, that are being applied to real environments [18].

Summarization module. The goal of this module is to produce summaries from a certain number of texts of the same topic. This is an adaptation of the work presented in [19], with the difference that the module is designed to be language-independent. After a preprocessing stage that uses the NLP module, the process collects all the recognized named entities and keywords of the text, using the well-known Term Frequency-Inverse Document Frequency (TF-IDF) algorithm [20] for this second task. Once each text has been preprocessed, the module determines the most relevant, non-repeated sentences using a double clustering approach: clustering by similarity and clustering by keywords. After that, the process sorts the sentences and determines if the

[1] http://nlp.lsi.upc.edu/freeling/

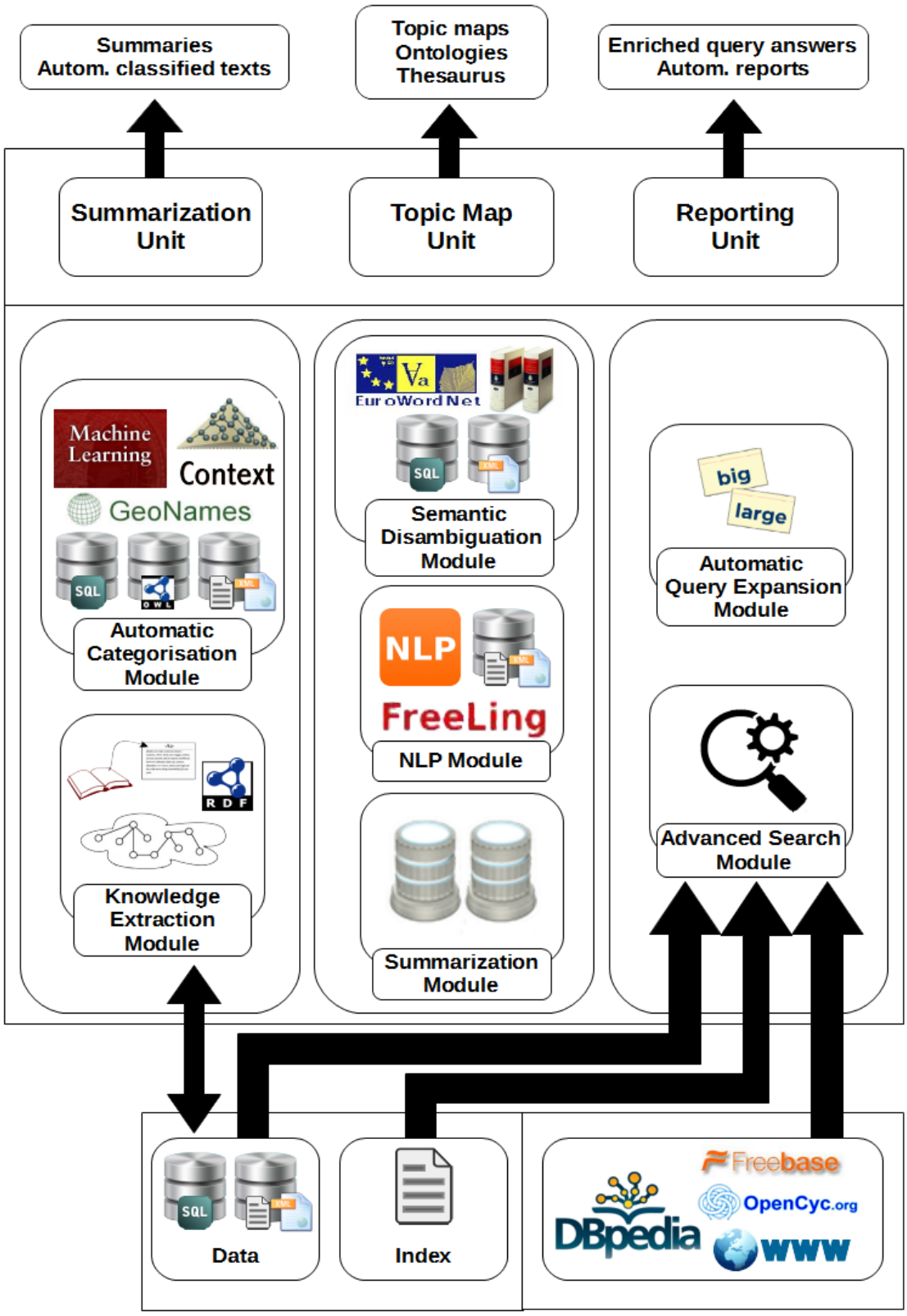

Fig. 1. Detailed architecture of Hypatia

ones with the highest scores compose a correct summary. Finally, the module implements an adaptation of the approach presented in [21] to perform a sentence simplification. This final step leads to a more compressed summary that contains only the main content of the sentences, with no complementary information.

This module is used in other libraries of Hypatia to create more specific and graphic representations of the information.

Semantic disambiguation module. This module is an unsupervised disambiguation engine designed to carry out the disambiguation of word meanings. The main task of this engine is to find the most accurate meaning of a term in a certain context. It includes a set of methods and techniques that compute the semantic relation between words. For this purpose, it uses a set of dictionaries with semantic information of words and a lexical thesaurus to find the possible sets of synonyms for a term. The lexical thesaurus used has been EuroWordNet [22], which is a multilingual database for several European languages, such as Spanish, French, Italian, and German. Each language designs its own set of words by structuring it as WordNet, using sets of synonyms with basic relations among them.

Specifically, this module combines measurement techniques based on conceptual trees and based on the content of the information, taking advantage of EuroWordNet. It uses a given word as input and a set of lists of meanings of that word provided from the aforementioned lexical resources, and then it computes the degree of semantic similarity between the term and each possible list of meanings. The remaining libraries in Hypatia will use this module when a disambiguation is required.

B. Preprocess Unit

Before the users can work with Hypatia, there is a preprocessing stage in which the expert system feeds itself with available information in order to classify it. This information is extracted from internal sources of the company defined by direct connections and/or web services. Once all the information has been classified and optionally stored inside an internal database, Hypatia indexes it using an inverted index which will provide a quick access to the stored data.

Categorization module. If there is an available thesaurus in the expert system, this module uses it to perform a classification of the data. The detailed pipeline of the process, which is an approach that combines NLP and machine learning, can be found in [23]. This module has a preprocessing stage in which it uses the NLP module to perform a morphological analysis of the input data, along with the named entities that it may contain. After that, it applies the previously mentioned TF-IDF algorithm to extract keywords from the data. Once the preprocessing stage has finished, this module applies different classifiers in ascending order of complexity to perform an automatic classification. Firstly, it takes advantage of the attributes of each data item to make a simple rule-based tagging. Secondly, it uses machine learning techniques [24] to determine whether certain data belong to a particular topic, and classify them according to the corresponding thesauri. Then, the module uses the named entities and applies a geographical classifier to get the geographic location the data is referring to (e.g., the location where a certain event described in the document occurs). Finally, a detailed labeling is performed using an ontological classifier, where all the keywords and named entities in the knowledge base are queried. If a positive response is obtained, the main related concepts of the response can be used to label the text.

Knowledge extraction module. If the data have no classification or the users want to improve the existing one, Hypatia can take advantage of the keywords and named entities obtained from the previous module to create new classification lists, thesaurus, and/or simple ontologies. To achieve that, the system follows the pipeline stated in [25] to create a topic map [26] from which it can generate the rest of the mentioned outputs. Considering the previously created summaries as the input of the process, this module analyzes the summary sentences with the NLP module in order to identify the function of each word in the sentence. After that, it considers the possible candidates to be elements in the topic map: the subjects and direct objects will be the topics and the complements respectively, and the verbs will be the associations between a topic and a complement. Once the topic map has been built, the module simplifies it by deleting the existing redundancies, incompatible associations, an ambiguities. Then, the simplified topic map is automatically evaluated and validated according to the standards of topic maps. Finally, the obtained topic map can be merged with others that belong to the same topic. The final topic map will be used as a basis for the conversion to the other outputs considered, for instance, using TM-Builder, an ontology builder based on topic maps.

C. Search Unit

After the user has stated the search query and the filters he/she may want, Hypatia performs the search using the modules inside this unit to obtain the information from its available sources.

Automatic query expansion module. This is a library that performs an automatic query expansion, based on NLP and semantics to improve information retrieval from relational databases repositories. This module has been implemented based on the methodology called SQX-Lib [27]. This module exploits semantic techniques to improve the quality of the keyword-based search process on relational repositories and it is focused on expanding the scope of the searches, and fine-tune them by taking advantage of the Named Entities (NE) [28] present in the query string. The aim of the process is to expand a query semantically to obtain a set of data that not only matches the keywords introduced in the search, but that it also contains data that include words that share the lemmas of the keywords and a similar meaning.

Semantic Search module. Once the final keywords and filters have been determined to specify the scope of the search, this module performs the search over the different sources Hypatia has access to. The expert system not only looks in its internal data repositories, but it also searchs on the Web. These external sources can be, for instance, search engines, social networks, or Linked Data [1]. Each of the different sources are treated using different strategies that depend on the data structure to extract. Regardless the origin of the information, the data is stored later in the same way for a better organized manipulation.

D. Displaying the Information

Hypatia has two main interfaces. The first one is centered around the system configuration and advanced options for the documentation departments, while the other one is the user interface. The user interface consists only in a search box along with some basic filter options to narrow the scope of the search and the outputs that are going to be shown. If the user inputs some ambiguous keywords, the system has to perform semantic disambiguation [29] in order to determine the possible senses that the user might refer to. Once the main senses have been determined, Hypatia can try to guess the desired search, or let the user to manually choose from the available options.

When all the output has been extracted and collected, Hypatia is ready to display the search results. As a result of the semantic search and the work performed by the previous modules, the data has semantic links that refer to complementary data. The expert system benefits from this to improve data navigation and data aggregation. To achieve that, the interface units use all the previously defined modules.

Summarization unit. Each set of data about a certain common aspect of the search is combined into a single unit of data, using the summarization module, to get the main ideas of that aspect. This unit is part of the aggregation main process, whose components are formed using all the extracted data. For instance, if Hypatia finds multiple results about the history of the marathon of New York, it will create a summary of it before showing it along with the rest of the report.

Topic map unit. If requested, Hypatia presents an option that allows to create a topic map about the search. In that case, it uses a topic map drawing tool like ONTOPIA to graphically represent the information, as the tool can draw topic maps from an XTM file, which is one of the outputs of the knowledge extraction module.

Aggregation unit. After the user has selected the desired search request in the interface, Hypatia proposes a list of reports that may be interesting for the user. These reports are elaborated from a set of predefined templates that define the presentation of the information of some types of results (people, companies, events, etc.). Continuing with the marathon of New York example, this result would fall inside the "event" category, and all the extracted information would be structured using that template.

These templates also drive some of the sources from where Hypatia extracts the information. The expert system extracts the information even if the sources are unstructured. In this case, it uses the NLP module and ad-hoc rules to identify patterns of information that may be interesting to extract. When all the information has been collected and summarized using the summarization module and the summarization unit, the final report is composed.

IV. Evaluation

The process of verifying that an expert system is accurate and reliable has two distinct components: checking that the knowledge base contains all the necessary information and verifying that the program can interpret and apply this information correctly to provide accurate information to the user. Some studies, such as [30], propose tools to check knowledge bases, but this kind of evaluation does not apply to Hypatia because it is not the typical rule-based system.

Therefore, to evaluate the system, a study with final users has been performed. This study consists on an opinion survey among a group of 100 workers of several Spanish media companies (Grupo Heraldo, Diario de Navarra and CARTV, among others). The users have been classified in two groups: 20 are expert documentation workers, and the other 80 are final users without advanced knowledge on documentation tasks. In this survey, they use a prototype that introduces the capabilities of the tool before being asked about its main features: semantic disambiguation, automatic query expansion, semantic filters and semantic relations between elements, automatic summaries, elaboration of personalized dossiers and reports, obtention of indexes and thesauri, the possibility of simultaneously search both within the internal database and in the Web, topic map generation, and the interface itself of the expert system. For each of the characteristics of the expert system, the user feedback is picked regarding the relevance of the feature *(Relev)*, its usefulness in daily work *(Util)*, the proposed manageability *(Man)*, and finally the novelty *(New)* of this functionality for the user in such tools. All the features are evaluated from 0 to 10. The results of the survey are shown in tables I and II.

TABLE I
SURVEY AMONG THE PROFESSIONAL USERS

Feature	Relev	Util	Man	New	Total
Disambiguation	10	10	9	6	8.75
Query Expansion	6	6	8	6	6.5
In/Out Searches	9	10	9	8	9
Filters & Relations	9	9	8	6	8
Autom. Summaries	8	9	9	9	8.75
Autom. Topic Maps	10	8	8	10	9
Autom. Dossiers	9	9	7	10	8.75
Indexes & Thesauri	9	7	8	10	8.5
Interface	10	10	8	8	9

TABLE II
SURVEY AMONG THE FINAL USERS

Feature	Relev	Util	Man	New	Total
Disambiguation	9	8	8	9	8.5
Query Expansion	9	9	7	8	8.25
In/Out Searches	8	9	8	9	8.5
Filters & Relations	9	8	9	9	8.75
Autom. Summaries	9	9	9	9	9
Autom. Topic Maps	6	5	9	10	7.5
Autom. Dossiers	9	9	9	10	9.25
Indexes & Thesauri	7	5	6	10	7
Interface	10	10	7	9	9

The functionalities that are best appreciated by all the users are the disambiguation, semantic filters and semantic relations between elements, automatic summarization, and dossier generation, with an average of 8-9 points. Professional users consider very interesting everything related to obtaining knowledge, whereas final users did not understand very well these options. Moreover, final users were delighted with the automatic query expansion, while documentalists did not find

it especially useful; this seems to be due to the different search style of each group of users and the different types of information that they want to find. Documentalists usually prefer to control the terms used in a query by themselves, but they might use query expansion as an optional feature.

V. Conclusions and Future Work

This work has studied the possibilities of Hypatia, an expert system aimed at professional documentation departments and end users of all kinds of documental information systems. The main idea of this system is to help users to search, extract, and organize the text-based data stored in a company's documental database, with the possibility of enriching the information retrieved with the help of semantic tools that query public external sources located in the Web, such as Linked Data. Hypatia integrates several types of existing tools and introduces some ideas that improve its functionalities and performance, such as the use of automatic dossiers, the use of automatic query expansion, or the semantic combination of different sources to retrieve the search results.

An interface prototype has been designed and developed to test the behavior of the system with final users. Moreover, a survey with users of real companies has been conducted to analyze their feedback. This last has been performed using a twofold approach: on the one hand from a documental expert perspective, and on the other hand from a non-professional user point of view. The results of this study are very promising, and highlight the interest of the system. According to the authors knowledge, there is no other similar expert system offering the range of functionalities that Hypatia covers.

Since the survey was conducted with employees of media companies, the next step is to further test Hypatia in other activity sectors in order to verify the generality of the solution. Besides, there is work in progress to improve the creation of automatic reports, which is still in an experimental phase.

Acknowledgment

This research work has been supported by the CICYT project TIN2010-21387-C02-02 and DGA-FSE. The authors want to sincerely thank Maria G. Buey and Sandra Escudero for their useful suggestions.

References

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data – the story so far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[2] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[3] A. F. Smeaton, "Using NLP or NLP resources for information retrieval tasks," in *Natural language information retrieval*. Springer, 1999, pp. 99–111.

[4] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[5] R. B. Mishra and S. Kumar, "Semantic web reasoners and languages," *Artificial Intelligence Review*, vol. 35, no. 4, pp. 339–368, 2011.

[6] A. Gilchrist, "Thesauri, taxonomies and ontologies–an etymological note," *Journal of Documentation*, vol. 59, no. 1, pp. 7–18, 2003.

[7] M. Atkinson and E. Van der Goot, "Near real time information mining in multilingual news," in *18th international conference on World Wide Web*. ACM, 2009, pp. 1153–1154.

[8] J. X. Yu, L. Qin, and L. Chang, "Keyword search in databases," *Synthesis Lectures on Data Management*, vol. 1, no. 1, 2009.

[9] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[10] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM*, vol. 16, no. 2, pp. 264–285, 1969.

[11] C.-Y. Lin and E. Hovy, "Identifying topics by position," in *5th Applied Natural Language Processing Conference*. Association for Computational Linguistics, 1997, pp. 283–290.

[12] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[13] P. Jaccard, *Nouvelles recherches sur la distribution florale*. Impr. Réunies, 1908.

[14] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1998, pp. 335–336.

[15] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.

[16] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, "Multi-document summarization by sentence extraction," in *2000 NAACL-ANLP Workshop on Automatic Summarization*, vol. 4. Association for Computational Linguistics, 2000, pp. 40–48.

[17] L. Serrano, M. Bouzid, T. Charnois, S. Brunessaux, and B. Grilheres, "Events extraction and aggregation for open source intelligence: From text to knowledge," in *25th International Conference on Tools with Artificial Intelligence*. IEEE, 2013, pp. 518–523.

[18] A. L. Garrido, M. G. Buey, S. Escudero, A. Peiro, S. Ilarri, and E. Mena, "The GENIE project - a semantic pipeline for automatic document categorisation," in *10th International Conference on Web Information Systems and Technologies*. SCITEPRESS, 2014.

[19] S. B. Silveira and A. Branco, "Extracting multi-document summaries with a double clustering approach," in *Natural Language Processing and Information Systems*. Springer, 2012, pp. 70–81.

[20] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[21] S. B. Silveira and A. Branco, "Compressing multi-document summaries through sentence simplification," in *5th International Conference on Agents and Artificial Intelligence*, 2013, pp. 111–119.

[22] P. Vossen, *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Boston, 1998.

[23] A. L. Garrido, O. Gomez, S. Ilarri, and E. Mena, "Nass: news annotation semantic system," in *23rd International Conference on Tools with Artificial Intelligence*. IEEE, 2011, pp. 904–905.

[24] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[25] A. L. Garrido, M. G. Buey, S. Escudero, S. Ilarri, E. Mena, and S. B. Silveira, "TM-gen: A topic map generator from text documents," in *25th International Conference on Tools with Artificial Intelligence*. IEEE Computer Society, ISBN 978-1-4799-2971-9, ISSN 1082-3409, November 2013, pp. 735–740.

[26] S. Pepper and G. Moore, "Xml topic maps (xtm) 1.0-topicmaps.org specification," *TopicMaps.Org Authoring Group*, vol. 23, no. 08, p. 2003, 2001.

[27] M. G. Buey, A. L. Garrido, S. Escudero, R. Trillo, S. Ilarri, and E. Mena, "Sqx-lib: Developing a semantic query expansion system in a media group," in *36th European Conference on Information Retrieval. Demo paper*. Springer Verlag LNCS, 2014.

[28] S. Sekine and E. Ranchhod, *Named Entities: Recognition, Classification and Use*. John Benjamins, 2009.

[29] R. Trillo, J. Gracia, M. Espinoza, and E. Mena, "Discovering the semantics of user keywords," *Journal of Universal Computer Science*, vol. 13, no. 12, pp. 1908–1935, 2007.

[30] M. Suwa, A. C. Scott, and E. H. Shortliffe, "An approach to verifying completeness and consistency in a rule-based expert system," *Ai Magazine*, vol. 3, no. 4, p. 16, 1982.