How ontology based information retrieval systems may benefit from lexical text analysis

Sylvie Ranwez, Benjamin Duthil, Mohameth-François Sy, Jacky Montmain, Patrick Augereau, Vincent Ranwez

How ontology based information retrieval systems may benefit from lexical text analysis

Sylvie Ranwez, Benjamin Duthil, Mohameth François Sy, Jacky Montmain, Patrick Augereau and Vincent Ranwez

1 Introduction

The exponential growth of available electronic data is almost useless without efficient tools to retrieve the right information at the right time. This is especially crucial with respect to decision making (e.g. for politicians), innovative development (e.g. for scientists and industrial stakeholders) and economic development (e.g. for market or competitive analyses). It is now widely acknowledged that information retrieval systems (IRSs in short) need to take semantics into account to enhance the use of available information. However, there is still a gap between amounts of relevant optimized IRS information that can be accessed on the one hand, and users' ability to grasp and process a handful of relevant data at once on the other. Even though Semantic Web technologies and ontologies are now widespread and accepted, they are hampered by the fact that they cover few aspects that a document deals with: this is known as the semantic gap issue. They should thus be jointly used with terminological or lexical approaches to enrich document description.

This chapter starts with a survey on semantic based methodologies designed to efficiently retrieve and exploit information. Hybrid approaches including lexical analysis are then discussed. Terminology based lexical approaches are tailored to open contexts to deal with heterogeneous and unstructured data, while other taxonomy or ontology based approaches, are semantically richer but require formal

S. Ranwez, B. Duthil, M-F. Sy, J. Montmain
LGI2P research center from École des Mines d'Alès, Parc scientifique G. Besse, F-30035 Nîmes Cedex 1, France, firstname.lastname@mines-ales.fr

V. Ranwez
SupAgro Montpellier (UMR AGAP), 2 place Pierre Viala, F-34060 Montpellier Cedex 1, France, ranwez@supagro.inra.fr

P. Augereau
IRCM, Institut de Recherche en Cancérologie de Montpellier Inserm U896 and Université Montpellier1, CRLC Val d'Aurelle Paul Lamarque, F-34298 Montpellier, France, patrick.augereau@inserm.fr

knowledge representation of the studied domain and conceptual indexing. While these latter are often implemented at the document level, automatic terminology indexing allows fine-grained descriptions at the sentence level. Hence, there is a continuum of solutions from terminology to ontology based IRSs. These approaches are often seen as concurrent and exclusive, but this chapter asserts that their advantages may be efficiently combined in a hybrid solution built upon domain ontology. The original approach presented here benefits from both lexical and ontological document description, and combines them in a software architecture dedicated to information retrieval in specific domains. Relevant documents are first identified via their conceptual indexing based on domain ontology, and then each document is segmented to highlight text fragments that deal with users' information needs. The system thus specifies why these documents have been chosen and facilitates end-user information gathering.

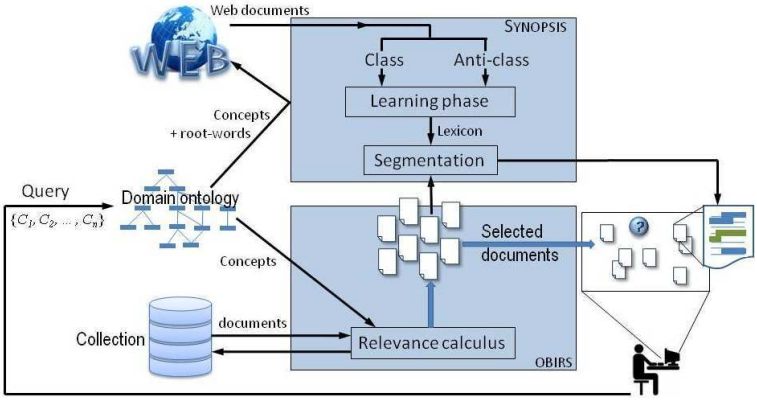

Fig. 1 *Overview of our CoLexIR approach*

Section 2 reviews different IR strategies and highlights the performance obtained using conceptual approaches, including OBIRS, an ontological based information retrieval system [55] which is more detailed as it serves as a basis for our hybrid IRS. Section 3 reviews the foundations of concept identification through lexical analysis and details the different phases of text segmentation that are implemented within the *Synopsis* approach [13]. Then Section 4 proposes an architecture to combine lexical analysis with a conceptual approach, according to the overall *CoLexIR* (Conceptual and Lexical Information Retrieval) view given in figure 1. The pros and cons are discussed, particularly the complementarity of both approaches is underlined and we show how their limits may be overcome by this combination. Two kinds of evaluation of this environment are proposed in Section 5. The first one relies on public benchmarks using some publicly available document collections. The second involves expert evaluations to assess the relevance and man-machine interactions in our system, using a set of BMC cancer publications as corpus. This latter

and its indexing by medical subject headings (MeSH[1]) concepts are freely accessible via PubMed (biomedical literature from the National Center for Biotechnology Information). These tests focus especially on documents preview that facilitates and speeds up bibliographic research by pinpointing relevant sentences from relevant documents. Some perspectives are finally given, particularly concerning the possibility of automatic indexing approaches in closed contexts (data warehouse containing similar documents indexed using ontological concepts) and their extension to open contexts (the Web containing heterogeneous and poorly indexed documents).

2 Information retrieval: keywords vs. concepts

The main task of an information retrieval system (IRS) is to select information which is likely to meet user needs, expressed as queries. Three processes are usually implemented in IRSs to fulfil this task [34]: i) an indexing process which aims to provide a representation as compact and expressive as possible of resources (textual, multimedia documents) and queries; ii) a matching process for selecting relevant resources w.r.t. to a query; iii) a query reformulation process that typically occurs between the two previous points. IRS may thus be seen as a function that maps a query Q (user information need) from a query set H_Q to a set of m documents within collection D of all indexed documents (also called corpus).

$$IR : H_Q \rightarrow D^m \tag{1}$$

Document and query indexing models (singleton or complex structure) and query-document matching strategies (strongly dependent on the indexing model) are generally sufficient to characterize and identify information retrieval models. Many IRSs have been proposed in the literature, depending on the relevance estimation process: i) Boolean models; ii) vector models, which represent documents and queries as weighted vectors (with distinct indexing units as base) and their suitability is estimated using vector distances; iii) probabilistic models, which can be subdivided into two categories: those (the most classical ones) that consider relevance as a binary variable whose probability is estimated; and language models where relevance is the probability of deducing queries from document models.

In [14], IRSs relevance models are presented as aggregation processes and formalized as a quadruplet (D, F, ε, P), where:

- D: is the set of documents to be evaluated and ordered w.r.t their accuracy to a user's query;
- F: is a family of criteria to assess a document's relevancy;
- ε: is the set of performance vectors; an elementary relevancy degree is associated with each document w.r.t a criterion, then the overall relevancy of a document is modeled by the vector of elementary relevancy degrees w.r.t all the criteria;

[1] http://www.ncbi.nlm.nih.gov/mesh

- P: is an aggregation function that accounts for the order relationship induced by the elementary relevancy degrees.

Indexing plays a key role because it provides content description of resources, allowing search tools to match them with user queries. Depending on indexing methods, IRSs are historically classified in two categories [20]: keyword-based IRSs, also called *syntactic* search systems, and conceptual IRSs, known as *semantic* search systems.

2.1 Keyword-based IRSs

Keyword-based IRSs often represent documents and queries as a bag-of-weighted-words or multiwords (phrase). This representation is obtained through document lexical analysis within collections, that summarize document contents by a set of lexical units [15]. A keyword-based IRS relevance process may rely on an exact match, an approximate match, or a string distance between words within documents and query indexing. Hence, when a query is submitted, these systems will retrieve documents indexed by exact query keywords or some of their lexical variations (e.g. *tumorous* instead of *tumor*). Unfortunately, they miss documents having query keyword synonyms in their indexing (e.g. *carcinoma* instead of *tumor*) [20] [4] [18]. This so-called the synonymy problem is the most common shortcoming, but keyword-based IRSs also fail to consider various kinds of semantic relationship between words (hyponyms, hypernyms). They are hampered by polysemous problems due to language ambiguity [3] [18]. Indeed, a word may have several meanings depending on the usage context (e.g. *cancer* as astrological sign or as illness). A syntactic search engine will retrieve a document when its indexing contains a query keyword, even if the meaning of the word within the document differs from what the user had in mind. All of these issues account for the lack of precision of keyword-based information retrieval systems, which is a well known problem [53].

Two solutions have been proposed to solve the above syntactic search limitations. Both of them involve improving indexing by introducing some semantics:

- Structuring lexical units (e.g. noun phrases) extracted from documents using some kinds of relationship (synonymy, subsomption, etc.). This is possible using natural language processing and machine learning techniques. This strategy may be seen as a first step towards interfacing ontologies and lexical resources since structuring of the latter involves ontological principles [42]. This approach is still, nevertheless, syntactic since the semantics remain implicit.
- Use of conceptual resources to represent document content based on their meaning rather than their words. These resources may be arranged from less formal ones (thesaurus with strong lexical compounds: WordNet or UMLS) to more formal ones (e.g. Gene Ontology). They can also be general or domain specific. Extraction techniques are needed to make use of such term meaning or concept for indexing purposes. These techniques may be manual or automatic[58], but

this topic is beyond the scope of this chapter. The corresponding term concept matching strategy leads to an ambiguity problem, e.g. a term appearing within more than one concept lexical compound, and thus disambiguation techniques are needed [19] [3].

These two strategies lead to different indexing units characterized by their granularity: from a lower level (lexical units such as words, noun phrases) to a higher level (conceptual units). The next section reviews and discusses the foundations of conceptual based IRSs.

2.2 Conceptual IRSs

As seen above, conceptual resources such as ontologies are used within the IR community to overcome some keyword-based system limitations. Conceptual IRSs are based on the assumption that document contents are better described by conceptual abstractions of real word entities than by lexical relationships that may be found within it or dictionaries [3] [11]. A cognitive view of the world is thus considered in such systems. The emergence of domain ontologies, boosted by the development of the Semantic Web (in its infrastructure and content), has led to an increase in conceptual IRSs. In these systems, ontology based concepts are used as pivot language for indexing documents and expressing queries. Such conceptual description of the word may also be used as a semantic guide while visualizing documents or data. Ontology also provides conceptual space in which metrics (semantic similarities or distances) can be deployed to implement the relevance calculus process in IRSs. According to [52], a domain ontology O can be formally defined as follows:

Definition 2.1 *$O := \{C, R, H_C, H_R, Rel, A\}$, where C and R are respectively a set of concepts and a set of nontaxonomic relations. H_C is a heterarchy of concepts with multiple inheritance. H_R is a heterarchy of relations. $Rel : R \longrightarrow C \times C$ defines nontaxonomic relations between concepts, while A is a set of logical axioms.*

2.2.1 Conceptual indexing

It is necessary to distinguish between conceptual and semantic indexing. Conceptual indexing comes from the IR community and relies on concept hierarchy or domain ontology (e.g. the ontology for biomedical investigation: MeSH), where documents are associated with a bag-of-concepts describing their contents. Semantic indexing comes from the Semantic Web community, where metadata are added to a knowledge database to characterize documents (resources). Semantic indexing is also called annotation within the Semantic Web community. Hereafter, domain ontology concepts are used to represent documents as conceptual indexing.

2.2.2 Query formulation

Even some concept-based IRSs allow users to express their queries in natural language. They include modules that extract information from these queries in order to transform them into a set of concepts that may be organized in different data structures, such as vectors [23] [11], trees [3], semantic networks, conceptual graphs [39] or bag-of-concepts. Other systems such as Textpresso [37] allow concept selection, while sometimes combining concepts using logical operators. In OBIRS [55], users are asked to choose among a set of proposed concepts they are supposed to be familiar with (assisted with automatic completion) in order to directly set their queries. The OBIRS interface also allows selection of parameter value that determines whether the query should be rather considered as an AND, OR, or something in between.

2.2.3 An example of concept-based IRS: OBIRS

[55] proposes an ontological-based information retrieval system (OBIRS) using semantic proximities and aggregation operators to assess document adequacy w.r.t a user query. Since OBIRS methodological details and validation protocols are available in [55], it is only outlined in this section.

An original multi-level score aggregation to assess the relevance of documents based on semantic proximity.

OBIRS allows assisted query formulation based on domain ontology concepts and implements a relevance model using semantic proximities. The proposed relevance score computation (also called retrieval status value [RSV]) consists of three stages of the aggregation process:

- The first stage computes a simple and intuitive similarity measure (denoted π) between two concepts of the ontology O. Several semantic proximity measures may be used here, that can be based on calculation of the shortest path, on the use of the information content (IC) [45] [31] or on set based measures [44]. In order to favor user interactions, concept proximities must be intuitive (so that the end-user can easily interpret them) and fast enough to compute (to ensure that the IRS remains efficient even in case of large ontologies). By default, OBIRS relies on Lin's proximity for this step [31].
- Then a proximity measure is computed between each concept of the query and a document indexing. Where d_i denotes the i^{th} element of the list $C(d)$ of concepts indexing a document d, the similarity between a concept Q_t of a user query Q ($t = 1..|Q|$) and d is defined as:

$$\pi(Q_t, d) = \max_{1 \leq i \leq |C(d)|} \pi(Q_t, d_i) \tag{2}$$

- Finally, the relevance score of a document w.r.t a query is assessed using the family of aggregation operators proposed by Yager [12]. Each query concept is

considered as a criterion to be satisfied and corpus documents as alternatives. The assessment of such alternatives with regard to the criteria is given by:

$$RSV(Q,d) = \left(\frac{\left(\sum_{t=1}^{|Q|} p_t . \pi(Q_t, d)^q \right)}{|Q|} \right)^{\frac{1}{q}} , q \in \Re, \sum_{t=1}^{|Q|} p_t = 1 \quad (3)$$

This aggregation model takes into account the user model preference about the kind of aggregation that has to be introduced to compute the overall relevance of a document w.r.t his/her query. When the above weighted operators's family is used, the user has to fit both q parameter and the p_t weights distribution upon the query terms. The weights characterize the relative importance granted to each of the query terms in the preference model, whereas the q parameter sets the extent to which the simultaneous satisfaction of all criteria is required to assign a high RSV score to a document. Indeed, in equation 3, when q has very small values ($-\infty$) the query tends to be conjunctive (aggregation involves the MIN operator) whereas when q gets close to $+\infty$, the query tends to be disjunctive (aggregation involves the MAX operator).

This last stage synthesizes document relevance and users' preferences and ranks the collection of retrieved documents according to their RSV. The aggregation model enables restitution of the contribution of each query concept to the overall relevance of a document. Hence it provides our system with explanatory functions that facilitate man-machine interaction and assists end-users in iterating their query.

2.2.4 Concept-based IRS issues

Complete conceptual indexing is hard to achieve in realistic collections. The reasons are twofold: firstly, domain ontologies may be hampered by weak coverage of all content aspects of the documents [4] and secondly, high quality indexing requires human expertise and is thus a tedious task. This is known as the *semantic gap* issue. Indeed, automatic or semi-automatic indexing techniques cannot always extract all significant document concepts. In order to increase ontology coverage and improve both document and user query indexing within conceptual based IRSs, lexical components could be added to the ontology, as detailed in the next section.

2.3 Hybrid ontology based information retrieval system

Hybrid IRSs have been designed to take both keyword based and conceptual based indexing units into account. We propose the following definition for an hybrid ontology based information retrieval system:

Definition 2.2 *An ontology based information retrieval system is called* hybrid *when it manages document indexes of different granularities (ontology based and keyword based) and levels (document level and passage level descriptions) during indexing and matching processes and/or during the result presentation stage.*

- *document index of different granularity:* ontology based and keyword based descriptions may be jointly used within *hybrid* ontology based IRSs, when the indexing process failed in attaching some keywords from documents, called instances, to concepts from the used ontology. In this case, the index may include a description that contains these keywords as well as a conceptual description that contains the identified ontology concepts. These latter may be a set of ontology concepts which can be organized in a complex structure (tree, network,etc.). They may also be a set of RDF triples which represent both concept instances and their relations within documents. This leads to hybrid relevance models, which are discussed below.
- *document index of different levels:* indexing units in both keyword based and ontology based IRSs may be related to the whole document (*document level*) or to some of its parts (*passage level*). Characterization of the document parts allows passage retrieval, which is suitable for multi-topic documents. In this case, passages are considered independently, indexed with concepts or keywords, and treated as documents. But characterization of document parts may also be used to give some hints about the document selection. Some explanations may be given to the user to justify, at the document level, the result selection. When the concepts occurring in different document passages are known, it is possible to segment texts and highlight passages that deal with user query concepts. This provides search engine users with insight into the IRS results, especially in biomedical literature [22][32]. A passage is not necessarily a paragraph within a document but any continuous subset (portion) of texts . This user interaction improvement will be discussed below.

Hybrid relevance model.

In hybrid relevance models, the two granularity document descriptions are considered separately since they do not describe the same viewpoint on the document. This assumption is based on the fact that a document keyword is used as an indexing unit only when information extraction tools failed at connecting it to a concept within the ontology. This independence assumption leads to hybrid IRSs that propose relevance models using two kinds of document/query suitability assessment: conceptual or semantic based and keyword based. A merged strategy of these two outputs is then applied. Three kinds of query are thus possible in such hybrid relevance models:

- fully semantic or conceptual queries (using only ontology concepts or relations);
- fully keyword queries (no semantic description of documents is available);
- mixed queries (both keyword and conceptual queries are available).

Many hybrid relevance models have been proposed in the literature. The authors in [4] propose *K-search*, an implementation of a hybrid search strategy. K-search com-

bines ontology based and keyword based search to support document retrieval (*ad hoc* retrieval) and knowledge retrieval (RDF triples retrieved using Sparql). Having two kinds of document descriptions (RDF triples with a link to their resources and keyword indexes), they define a hybrid relevance model as the combination of a keyword based model (e.g. using *Lucene*) and a semantic model (like *Sesame*) used independently. Keyword searches return a set $\triangle$ of documents. Semantic searches return a list of RDF triples associated with the documents they come from, and thus the set $\triangle P$ of these documents is built. The overall *K-search* results consist of the intersection of $\triangle$ and $\triangle P$.

[18] extends keyword search and defines three dimensions in a Cartesian space where semantics may help to fix some issues. This extension is called a *semantic continuum* Fig.2. Each dimension supplements the keyword search using semantic search when possible. The first dimension goes from natural language to formal language in order to solve keyword search polysemous and synonymy problems (word to concept). Considering this axis, a 0 coordinate means that the IRS considers words as indexing units whereas 1 means that conceptual units are used. When systems move through this axis, some words may not be mapped to a concept due to a lack of background knowledge (weak coverage: semantic gap). Authors propose to use both syntactic and semantic retrieval to overcome this drawback. This axis deals clearly with indexing granularity. The second axis ranges from words to phrases to overcome complex concept expression. This dimension deals with indexing structure. A 0 value on this axis corresponds to single indexing units (word or concept) while 1 represents complex ones. The last axis goes from string similarity to semantic similarity to achieve relatedness estimation of indexing units.

Organizing hybrid IRSs in such a 3D Cartesian space provides a simple and relatively intuitive characterization of these IRSs but may give rise to some limitations. Indeed, three dimensions are insufficient to fully describe IRSs not only because of the kind of indexing process they implement, but also because of the complexity of the relevance calculus and user behavior. Moreover, there is no proof of the independence of the chosen dimensions. Finally, a linear axis is not sufficient to represent an index complex structure since the information on how units are linked and organized are not taken into account.

Hybrid approach for user interaction improvement.

Information retrieval is often an iterative process where the user refines his/her query to focus on highly relevant documents. But this process implies that the end-user has a precise understanding of the results proposed by the search engine and that interaction techniques allow him/her to reformulate the query, select interesting documents and give some hints to the system about his/her application needs. Visualization techniques may thus be considered as key components of this process since they play a mediating role in this understanding. There are many specifications that characterize IRS result visualization interfaces but two of them are of particular interest:

- Cognitive aspects. In consideration of users' cognitive limits, it is important to: provide relevant document identification at a glance, enable users to focus on a

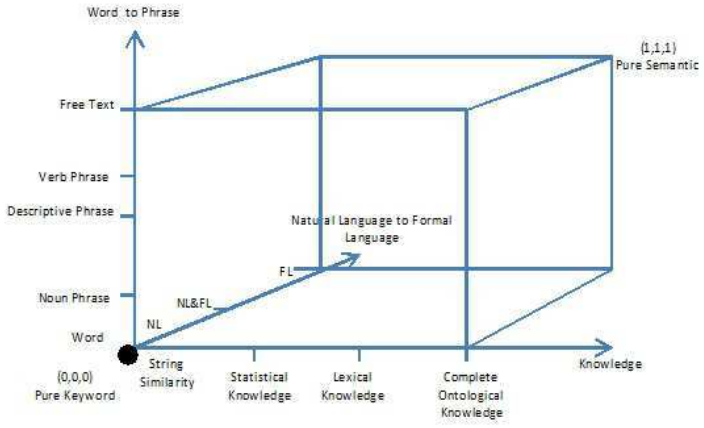

Fig. 2 *Semantic continuum*: *the classification of hybrid IRSs as proposed by* [18]

specific section of the visualization interface, and to intuitively understand any action results [56].

- Colors. Visual color scanning requires less time than visualizing words [10]. Colors may highlight some semantics and reflect the relative importance of displayed elements: e.g. green may denote the presence of a query term in a document indexing unit, while red may indicate the presence of another more specific related term.

Dimensionality is also an important feature of IRS visualization interfaces. However, most IRSs display results in a 2D space.

The simplest and most common way to display query results is in a list, where each item includes the retrieved document's title and its snipet with query terms highlighted (concept by label identification or words) in the document context. However, this type of presentation does not meet the above requirements. When passage level description is available, hybrid IRSs are able to show result explanations at the text level, thus synthesizing relevant information by highlighting relevant passages. Many such systems propose a range of result displays, from traditional document lists to passage visualization [32]. In *K-search* ([4]), retrieved documents are displayed in a list and document details are available in a separate panel when one of them is selected: keywords and RDF triples are thus highlighted. *K-search* also allows summarization of results using bi-dimensional graphs where different variables (e.g. retrieved document location) can be plotted. This graph is used to filter results. With the *Ontopassage* search engine [32], long and multitopic documents are fragmented as sets of passages and used as collection units. These passages are indexed using an ontology constructed from domain resources (e.g. relevant technical books of a domain). The system allows users to use different relevance models in the same query session (vector space model, probabilistic model). Users can switch from a traditional display mode (list of retrieved documents) to a passage display mode. In the latter, most relevant passages w.r.t. the user query of each retrieved document are

displayed. A small concept hierarchy is also displayed for each document, allowing users to explore related query concepts.

The relevance assessment model developed in OBIRS relies only on an aggregation model, as stated in 2.2.3. The relevances of retrieved documents using different IR models are not comparable. Therefore, we consider that allowing users to switch between these different models introduces confusion with IR visualization interfaces.

The ontology has to be supplemented with lexical resources so as to be able to identify document passages that are related to domain ontology concepts. Most domain ontology construction methods do not hold lexical information from which its concepts are taken. The formalisms used to represent an ontology, such as OWL, focus on intrinsic description of concepts, property classes and logical constraints on them. Many initiatives have been conducted to link ontology concepts to lexical information [54]. Interfacing techniques are needed in order to take both conceptual knowledge and lexical information in hybrid IRSs into account. [42] distinguishes three different approaches. The first one aims at structuring lexical resources using ontological principles without ontological category or relation. The second uses lexical information to enrich an ontology by adding lexical entries to the ontology (populating) [40] or by adding lexical information to concepts. Adding lexical entries to an ontology may increase the ontology size and coverage, whereas enriching ontology concepts with lexical information does not change the ontology structure even if the coverage is increased. The last way of interfacing ontology and lexical resources combines the two previous approaches.

[52] provides a definition of a lexical component of an ontology O:

Definition 2.3 *A lexical component L of an ontology O is defined as: $L := \{L_C, L_R, F, G\}$ where L_C,L_R are disjoint sets of lexical entries respectively related to concepts and relations; F,G provide correspondence between concepts and their lexical entries (between relations and their lexical entries, respectively).*

In our approach, we first aim to attach lexical information to ontology concepts and, second, to use such lexical information to determine passages within documents that deal with each query concept in the returned results. Our enrichment methodology therefore does not change the ontology structure. The next section proposes an implementation of the F correspondence function, thereby producing a lexicon for concepts and a thematic extraction process.

3 Concept identification through lexical analysis

Our interface between ontologies and lexical resources refers to the *enriching* option that has been previously described as "attaching lexical information to ontology concepts". This section details the basic notions of text segmentation, particularly to identify parts which deal with a specific concept in a document. Our approach is highly connected to the text partitioning process and thematic extraction process.

3.1 Related work

A text partitioning process is based on the analysis of thematic breakdowns in a document in order to subdivide the document into semantically homogeneous parts. These parts are considered as "text portions" (passages) which have very strong semantic coherence and are clearly disconnected from adjacent parts [48]. Thematic text segmentation may also be seen as a process of grouping basic units (words, sentences, paragraphs, etc.) in order to highlight local semantic coherence [28]. From a global standpoint, thematic structure search [35, 36] is a first crucial analysis step in many applications such as text alignment, text summarization, or information retrieval [2].

Among approaches described in the literature, two categories may be distinguished:

- *Lexical cohesion based approaches*. Several approaches measure this cohesion via term repetitions, semantic similarity, context vector entity repetition, word frequency models or word distance models. The re-occurrence of specific terms may indicate the presence of a common topic [1, 21, 26]. Lexical chains and their extension, the so-called weighted lexical links approach, are two identification techniques often used in a huge collection. The topic unigram language model is the most frequently used technique [41]. Most lexical cohesion based techniques are linear topic segmentation algorithms. These algorithms set boundaries inside a text at positions where a topic shift is identified. This process is performed in a (fixed size) sliding window. Lexical variation often results in dropping an employed similarity measure. Many methods use this process: TextTiling [21], C99 [6], Dotplotting [46], and Segmenter [26].
  There are also other statistical approaches that use the overall information in the text [25]. Text segmentation is based on analysis of the whole text, contrary to lexical cohesion based approaches that analyze a text on the fly. Malioutov [33] presents a graph-theoretic framework. The text is converted into a weighted undirected graph in which the nodes represent sentences and the edges quantify thematic relations between them. Text segmentation is performed by maximizing the similarity within each partition and minimizing dissimilarity across the partition [51]. [30] offers a statistical linear segmentation based on genetic algorithms.
- *Natural language processing techniques*. Linguistic methods introduce a set of specific rules that link words to each other (e.g. N-grams). These rules are dependent on the corpus. Linguistic methods still use external semantic information resources such as thesauri and ontologies. Resulting information from the association rules and from external semantic sources may then be combined through statistical techniques [38], which are highly dependent on available resources. Caillet proposes an automatic segmentation method based on term clustering [5]. This approach discovers the different themes in a text and extracts their related representative terms. [9] proposes an algorithm to recombine segments according to their content.

Note that segmentation approaches all have the same weakness: they do not allow precise identification of the themes (labeling) of a text portion, they only detect semantic breaks in a text without providing labels. To solve this labeling issue, some studies, based on text summary [17] and key phrase extraction approaches [24] identify text portions or key phrases according to their major theme [8]. Other methods focus on the identification of text portions related to the document title [29]. Most automatic text summary methods are based on a supervised learning process, that requires human intervention to set an adequate training corpus [57, 7]. [47] proposes an unsupervised method to extract key phrases in a summarization context.

Similar to segmentation methods, the approach presented in the following uses statistical information to identify, in a non-supervised context, text portions related to a given concept.

3.2 The "Synopsis" approach

The aim is to automatically identify all parts of a document that are related to the same concept. Knowledge extraction from textual data is usually a crucial step in the document indexing process. This section describes an adaptation of the *Synopsis* approach [13] involving tagging of text items according to predefined concepts (e.g. those expressed in the user query). For each concept, the *Synopsis* process starts by building a lexicon L containing a set of words that characterize it and a set of words that do not. This is performed by processing a significant number of documents that are downloaded through a Web search engine (e.g. *Google*). Then, based on the learned lexicon, *Synopsis* identifies text portions according to the given concepts.

This section describes the two main phases of this process: i) generation of the learning dataset and elaboration of concept lexicons (3.2.1) and ii) extraction of topics related to the concepts from textual data (3.2.2).

As our hybrid approach evaluation (see section 5) relies on *Cancer* related scientific publications, some vocabulary in this domain will be used hereafter to illustrate our approach. The scientific publications are indexed by the *MeSH* ontology concepts.

3.2.1 Concept characterization

As a start, lexicons related to some concepts in a domain have to be built. There are four steps in this process: acquisition of relevant corpus for each concept, significant words learning, representativity calculus for each of these words and lexicon elaboration.

Acquisition of relevant corpus. The first objective is to automatically build a training corpus for each concept of interest in a specific domain. For our purposes, these concepts are those in the user's query and the domain is *cancer*. A set of *root-words*

(also called *germs*) has to be attached to each concept. Here we rely on the *MeSH* ontology to automatically obtain n root-words, which are the label of the concept of interest and those of its hyponyms. For example, regarding the "*dna*" concept, the following root-words may be identified thanks to its label and its hyponyms ones: "*dna*", "*dna, z-form*", "*dna, satellite*", "*dna, intergenic*", "*dna, plant*"... These root-words are said to be discriminant for the concept C. For each root-word r related to a concept C, the *Synopsis* system, via a Web search engine, searches for 300 documents that contain both the root-word r and the name of the domain (e.g. "*dna, z-form*" and "*Cancer*" in our case). Together, these texts will form the *class* of C. This is the first part of the corpus associated with C.

Similarly, the system searches for 300 documents of the domain that do not include any root-words of the concept C. Together, these texts are called the *anti-class* of C. This set constitutes the second part of the corpus related to C. It obviously improves characterization of the concepts: a domain term that appears frequently in the class as well as in the anti-class for a concept is not discriminating (not representative) for this concept.

The class related to C thus contains $n * 300$ documents (where n is the number of root-words). Its anti-class contains 300 documents. The union of the class and the anti-class of C constitutes the corpus related to C. The second step involves searching any words significantly related to the root-words within these documents.

Significant word training. First of all, HTML tags, advertising and other noisy contents are removed from the documents of the corpus related to C. These documents are then transformed using a morpho-syntactic analyzer and lemmatization techniques [49]. This step identifies the representative (respectively non-representative) words for C. This is achieved by occurrence frequency analysis, assuming that the probability that a word characterizes a concept is proportional to its occurrence frequency in the immediate neighborhood of one of the concept's root-words. This occurrence frequency is computed over the whole corpus of concept C and is used to quantitatively assess the representativity Sc of a word W w.r.t. C. At the end of this step, lexicon L related to a concept C is formed with a set of words and their representativity w.r.t. C.

Two categories of words are distinguished: i.e. those prevailing in the class and those prevailing in the anti-class.

More formally, the words in the immediate neighborhood of a concept's root-word are first selected with a window $\mathscr{F}$ in a document doc:

$$\mathscr{F}(r, sz, doc) = \{w \in doc / d_{noun}(r, w) \leq sz\} \tag{4}$$

where r is the root-word, sz represents the window size and $d_{noun}(r, w)$ is the distance corresponding to the number of nouns (considered as meaningful words [27]) separating a word w from r in the document doc [13].

Representativity of words. It is now possible, for each word W of the corpus, to define its representativity in the class of the concept C. It is denoted $X(W)$ and is the sum of occurrences of a word W in a window $\mathscr{F}(r, sz, d)$ for all the root-words

of C and all the documents of the corpus. Note that for the anti-class, there is a single "root-word" which is the domain itself. The representativity in the anti-class is denoted $\overline{X}(W)$.

Lexicon elaboration. From the representativity of a word W in the class and in the anti-class, a score is established for this word using the following discrimination function [13]:

$$Sc(W,sz) = \frac{(X(W) - \overline{X}(W))^3}{(X(W) + \overline{X}(W))^2} \tag{5}$$

The cubic numerator function allows a signated discrimination: words of the domain that are non-representative of the concept get negative scores, while representative words of the concept get positive scores. The square denominator function allows a normalized score. It is now possible to build a concept-specific lexicon. It will include a list of scored words for a given concept.

3.2.2 Thematic extraction

Finally, this section explains how to use the achieved lexicon in a thematic segmentation process. A sliding window $\mathscr{F}'$ is introduced: it is successively centred on each occurrence of nouns in the document.

From lexicon L of a concept C, a score is computed for the sliding window $\mathscr{F}'$ in a document doc in the following way:

$$Score_{doc}(\mathscr{F}') = \sum_{W \in \mathscr{F}'} Sc(W,sz) \tag{6}$$

For a given document doc, the sliding window $\mathscr{F}'$ is said to match a concept C when its score is greater than a threshold. This threshold is dertermined through a sensitivity analysis. In a nutshell, its choice corresponds to different levels of granularity linked to the thematic structure of a document, i.e. the possible points of view of a reader.

4 Human accessibility enhanced at the crossroads of ontology and lexicology

The use of domain ontology semantics is known to improve IRS effectiveness. Therefore *OBIRS* starts with a domain ontology (used for collection indexing and query expression) and computes the document relevance score using semantic proximities and aggregation operators. Our 3-stage relevance model (which allows RSVs to be computed) integrates both the semantic expressiveness of the ontology based data structure and the end user's preferences. It has been implemented and a web-

based client is available[2]. Although users want IRSs to return good relevant documents at the top of a results list, to ensure fast grasp of relevant information, they also need explanations about why they have been chosen, and indications about more interesting document parts [22]. *OBIRS* has evolved into a hybrid IRS at presentation level according to the definition given in section 2.3. This hybrid system, called *CoLexIR*, includes *OBIRS* and *Synopsis*.

In *CoLexIR* visualization interface, retrieved documents are displayed in a semantic map and placed according to their relevance score w.r.t. the query represented as a probe (symbolized as a question mark). The result explanation focuses on both conceptual and passage levels. The higher the score, the closer the document is to the query probe in the semantic map. Each document is represented by a pictogram which details its match with the query. The contribution of each query concept to the overall score assessment is summed up in a histogram where a bar is associated with each concept Q_t of the query. This bar is colored depending on whether the closest (according to the chosen semantic similarity measure) concept of the document indexing is exactly Q_t (green), a hyponym (red) or a hypernym (blue) of Q_t. The bar is purple in other cases. The size of the bar associated with Q_t is proportional to the elementary relevance of the document w.r.t. Q_t (i.e. $\pi(Q_t, D)$). Moreover, a more precise analysis of document relevances may be required. Therefore, passages that deal with each query concept are identified by the segmentation process and highlighted at the text level. Double clicking on a document shows passages related to each query concept. These passages do not necessarily contain any query concept labels but rather terms that have been related to the concept lexicons in the segmentation step. By this way, users may see their query concepts instances within each document and also other concepts that the document deals with and that could be used to refine their information needs. Figure 3 shows an overview of the *CoLexIR* visualization interface.

5 *Evaluation: results on benchmarks and user feedbacks on a real case study*

Our system is validated using two kinds of evaluation: the first one rests on public benchmarks using some publicly available documents collections; the second one implies experts who both assess relevance and man machine interaction of our system. These tests especially focus on documents personalized preview.

[2] www.ontotoolkit.mines-ales.fr/ObirsClient/

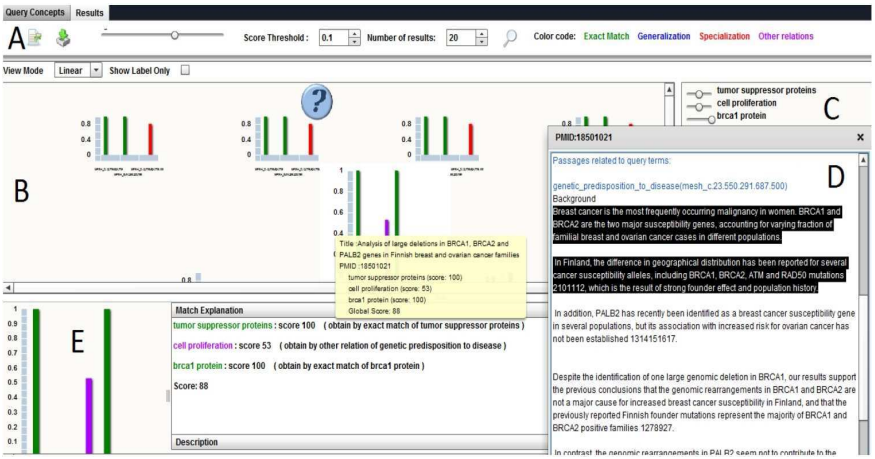

Fig. 3 *CoLexIR* interface displays selected document histograms in a semantic map according to their relevance scores w.r.t the query (symbolized by the question mark) (**B**). The query concepts and their weights are provided (**C**) as well as query parameters and color codes legend (**A**). Match explanation of a document is proposed as well as a link towards the whole document (**E**). Document passages related to the query concepts are available in a pop-up (**D**).

5.1 OBIRS results on an experimental campaign

The relevance model presented in this chapter, is experimentally validated following the Trec protocol. We use the MuchMore [3] collection which consists of 7823 medical paper abstracts and 25 queries with their relevance judgments. Documents and queries in that collection are indexed using MeSH concepts. The 1000 first retrieved documents have been considered and the precision of our system has been calculated at points 5, 10 and 20 as well as its mean average precision (MAP). To study the impact of IC based semantic similarity measures on OBIRS precision, we need to fix system parameters such as q value (set to 2.0), number of retrieved documents (1000) and RSV threshold (0.0, that means without filtering). Lin proximity measure [31] has been implemented and used for this experience. Our search strategy is also compared with Boolean search based on AND/OR operators. For all that measures, the intention driven (based on ontology not on corpus) evaluation of IC of a concept C from [50] is used:

$$\pi_{Lin}(C_1,C_2) = \frac{2.IC(MICA(C_1,C_2))}{IC(C_1)+IC(C_2)))}, IC(C) = 1 - \frac{log(hypo(C)+1)}{log(max_{con})} \quad (7)$$

where max_{con} is the number of concepts in the considered ontology, $hypo(C)$ the set of C hyponyms and $MICA(C_1,C_2)$ the most informative common ancestor (greater IC value) of C_1 and C_2. It should be noted that the IC value is 0 for the root

[3] http://muchmore.dfki.de

and 1 for leaves. The precision histogram in Figure 4 illustrates OBIRS precision values at some notable points. The obtained results show that our model (based on Lin) leads to better precision in considered points. Our method is also better than Boolean search strategy using “AND” or “OR” operators.

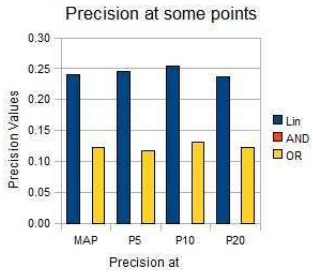

Fig. 4 *Comparison of CoLexIR (using Lin measure), AND, and OR search strategies on the Muchmore benchmark*

5.2 Experts experiments

Here we describe a biological case study in which the *CoLexIR* system is used to carry out a bibliographical study of proteins that could prevent cell proliferation induced by the BRCA1 protein. A first query of the three MeSH terms ”*tumor suppressor proteins*”, ”*cell proliferation*” and ”*brca1 protein*”, respectively weighted (100, 100, 100), was submitted to *CoLexIR*. *CoLexIR* detailed scoring of the retrieved documents enabled us to quickly determine that most of these documents did not often deal with the ”*brca1 protein*” MeSH term (low elementary score). A quick scan of *CoLexIR* excerpts of some of these retrieved articles confirmed that our query did not sufficiently stress our specific interest in BRCA1. We thus reformulated our query with adjusted weight, thus using ”*tumor suppressor proteins*” (50) + ”*cell proliferation*” (50) + ”*brca1 protein*” (100). This new formulation generated several relevant papers.

For most of the selected articles, the segmentation process highlighted some relevant pieces of information, w.r.t. query terms, that sometimes did not appear in the title or in the detailed abstract published by BMC cancer. For example, in [43], the *founder effect* noted in previous studies was not mentioned in the abstract, but retrieved by the segmentation process. The same was true for the fact that genomic rearrangement between BRCA1 and BRCA2 was not a major determining factor of breast cancer susceptibility in Finland, although this might be useful information for anyone interested in the genomic distribution of BRCA alleles in breast cancers. Similarly, in [16], several key results regarding *leukaemia* and *lymphoma* associated

genes were retrieved that were absent from the relatively long abstract of an article reporting the role of the BRCA1 gene in non-breast cancer. On the same lines, the excerpt concerning the interaction between BRCA1 and Fanconi proteins was valuable, and could provide researchers working in breast and immunological cancer fields with an opportunity to look for this interaction in either cancer type.

6 *Conclusion and perspectives*

Although lexical and conceptual approaches are mostly considered to be concurrent and exclusive strategies, hybrid IRSs can benefit from their complementarity to enhance information retrieval and presentation. Indeed, as stressed in the review proposed in this chapter, these two strategies are tailored to different kinds of system (open or closed), different granularity (document or sentences), and hybrid IRS aims to pull their strength. A review of these hybrid IRSs shows that most of them use different strategies to combine the document score of the two approaches so as to rank documents according to both view-points. They thus somehow still consider these two approaches as competitive solutions. We describe an alternative combination that we implement in a hybrid IRS dedicated to scientific articles retrieval. Relevant documents are retrieved via their conceptual indexing and then segmented to highlight passages that could be of particular interest for users.

The idea is to use each approach where it excels rather than to somehow average their points of view at each step of the search process. We thus propose to first use a conceptual model for document retrieval. The relevance of documents w.r.t. a query is then computed using both semantic similarity based on the conceptual model and users' preferences through a weight distribution over query concepts. Secondly, an explanation step, based on an original vizualisation system, helps users to gain insights into the results and facilitates interaction with the search engine for query reformulation. In addition to this relevance map, the user may require a more precise analysis of the document relevances. Each relevant document is thus segmented to highlight all the text portions related to the query concepts. The text portions do not necessarily contain any query concept labels but rather terms that have been related to the concept lexicons in the segmentation step. Users can have access to a more detailed analysis of relevance at a glance, while also identifying new relevant concepts to be able to more precisely express their information needs and then reformulate their queries.

The resulting *CoLexIR* system has been evaluated on the basis of its ability to retrieve relevant documents (using the MuchMore benchmark). A case study based on a corpus of BMC cancer papers highlights the usefulness of the *CoLexIR* functionalities and illustrates how its segmentation of retrieved papers allows users to rapidly identify relevant documents and grasp their key information (w.r.t. user needs) by reading sentences focused on the query terms. As expected, the main conclusions of the papers, as they appeared in the abstract, were actually selected by the seg-

mentation process. In addition, the excerpts helped place these conclusions in their context and retrieved additional relevant information scattered throughout the paper.

Excerpts selected by *CoLexIR* generally ranged from technical information (as found in "*material and methods*" sections or in figure legends) to general information (as found in abstract, introduction, discussion and conclusion sections). From a scientific standpoint, the technical information was generally of relatively low interest. The general approach of *CoLexIR* does not take the fact that BMC papers are strongly structured documents into account. It could be worth taking this information into account so as to enable end-users to select sections of scientific papers from which *CoLexIR* should extract excerpts. Further integration of lexical and conceptual approaches in *CoLexIR* could thus be beneficial. When scientific reviews do not impose the article to be structured using pre-defined sections, pre-processing of the corpus could be carried out in order to identify sections from which technical details are derived using a supervised lexical approach.

References

1. An, R.A., Morris, J., Hirstt, G.: Lexical cohesion computed by thesaural. Computational Linguistics **17**, 21–48 (1991)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. ACM Press/Addison-Wesley (1999)
3. Baziz, M., Boughanem, M., Pasi, G., Prade, H.: An information retrieval driven by ontology: from query to document expansion. In: RIAO (2007)
4. Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., Petrelli, D.: Hybrid search: effectively combining keywords and semantic searches. In: Proceedings of the 5th European semantic web conference on The semantic web: research and applications, ESWC'08, pp. 554–568. Springer-Verlag, Berlin, Heidelberg (2008)
5. Caillet, M., Pessiot, J.F., reza Amini, M., Gallinari, P.: Unsupervised learning with term clustering for thematic segmentation of texts. In: Proceedings of RIAO, pp. 648–656 (2004)
6. Choi, F.Y.Y.: Advances in domain independent linear text segmentation. In proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics **23**, 26–33 (2000)
7. Christensen, H., Kolluru, B., Gotoh, Y., Renals, S.: From text summarisation to style-specific summarisation for broadcast news. pp. 537–544 (2004)
8. Chuang, W.T., Yang, J.: Extracting sentence segments for text summarization: A machine learning approach. Proceedings of the 23 th ACM SIGIR pp. 152–159 (2000)
9. Clifton, C., Cooley, R., Rennie, J.: Topcat: Data mining for topic identification in a text corpus. In: In Proceedings of the 3rd European Conference of Principles and Practice of Knowledge Discovery in Databases (2002)
10. Cockburn, A., McKenzie, B.: 3D or not 3D? pp. 434–441 (2001)
11. Dragoni, M., Pereira, C.D.C., Tettamanzi, A.G.B.: An ontological representation of documents and queries for information retrieval systems. In: Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems - Volume Part II, IEA/AIE'10, pp. 555–564. Springer-Verlag, Berlin, Heidelberg (2010)
12. Dubois, D., Prade, H.: A review of fuzzy set aggregation connectives. Information Sciences **36**(1-2), 85–121 (1985)
13. Duthil, B., Trousset, F., Roche, M., Dray, G., Plantié, M., Montmain, J., Poncelet, P.: Towards an automatic characterization of criteria, DEXA '11. In: Proceedings of the 22nd International Conference on Database and Expert Systems Applications DEXA 2011, p. 457 (2011)

14. Farah, M., Vanderpooten, D.: A multiple criteria approach for information retrieval. In: SPIRE, pp. 242–254 (2006)
15. Fox, C.J.: Lexical analysis and stoplists. In: Information Retrieval: Data Structures & Algorithms, pp. 102–130 (1992)
16. Friedenson, B.: The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. BMC Cancer **7**, 152 (2007)
17. Gillick, D., Favre, B., Hakkani-tür, D.: The icsi summarization system at tac 2008. In: In Proceedings of the Text Analysis Conf. Workshop, pp. 801–815 (2008)
18. Giunchiglia, F., Kharkevich, U., Zaihrayeu, I.: Concept search. In: ESWC, pp. 429–444 (2009)
19. Gonzalo, J., Verdejo, F., Chugur, I., Cigarrin, J.: Indexing with WordNet synsets can improve text retrieval pp. 38—44 (1998)
20. Haav, H., Lubi, T.: A survey of concept-based information retrieval tools on the web. In: 5th East-European Conference, ADBIS 2001. Vilnius, Lithuania (2001)
21. Hearst, M.A.: Texttiling: segmenting text into multi-paragraph subtopic passages. ACM **23**, 33–64 (1997)
22. Hersh, W.: Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Briefings in Bioinformatics **6**(4), 344–356 (2005)
23. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G., Milios, E.: Information retrieval by semantic similarity. International Journal on Semantic Web and Information Systems **2**(3), 55–73 (2006)
24. Hulth, A., Megyesi, B.B.: A study on automatically extracted keywords in text categorization. In: In: Proceedings of 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (CoLing/ACL (2006)
25. Joris, D., Paul-Armand, V., Joris, V., Dirk, C., Joost, R.D.: Topic identification based on document coherence and spectral analysis. In: Information Sciences (2011)
26. Kan, M.Y., Klavans, J.L., McKeown, K.R.: Linear segmentation and segment significance. In: In Proceedings of the 6th International Workshop of Very Large Corpora, pp. 197–205 (1998)
27. Kleiber, G.: Noms propres et noms communs : un problème de dénomination. Meta pp. 567–589 (1996)
28. Kozima, H.: Text segmentation based on similarity between words. ACL pp. 286–288 (1993)
29. Kupiec, J., Pedersen, J., Chen, F.: A trainable document summarizer. Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval pp. 68–73 (1995)
30. Lamprier, S., Amghar, T., Levrat, B., Saubion, F.: Seggen: A genetic algorithm for linear text segmentation. IJCAI'07 pp. 1647–1652 (2007)
31. Lin, D.: An Information-Theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, pp. 296–304 (1998)
32. Lin, H.T., Chi, N.W., Hsieh, S.H.: A concept-based information retrieval approach for engineering domain-specific technical documents. Advanced Engineering Informatics (0) (2012)
33. Malioutov, I., Barzilay, R.: Minimum cut model for spoken lecture segmentation. In: In Proceedings of the Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006, pp. 25–32 (2006)
34. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
35. McDonald, D., Hsinchun, Chen: Using sentence-selection heuristics to rank text segments in txtractor. JCDL'02 pp. 28–35 (2002)
36. Misra, H., Yvon, F., Cappé, O., Jose, J.: Text segmentation: A topic modeling perspective. Information Processing and Management **In Press, Corrected Proof** (2011)
37. Müller, H., Kenny, E.E., Sternberg, P.W.: Textpresso: An Ontology-Based information retrieval and extraction system for biological literature. PLoS Biol **2**(11), e309 (2004)
38. Moens, M.F., De Busser, R.: Generic topic segmentation of document texts. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01, pp. 418–419. ACM, New York, NY, USA (2001)

39. Montes-y-Gómez, M., López-López, A., Gelbukh, A.F.: Information retrieval with conceptual graph matching. In: Proceedings of the 11th International Conference on Database and Expert Systems Applications, DEXA'00, pp. 312–321. Springer-Verlag, London, UK (2000)
40. Niles, I., Pease, A.: Towards a standard upper ontology. In: Proceedings of the international conference on Formal Ontology in Information Systems - FOIS '01, pp. 2–9. Ogunquit, Maine, USA (2001)
41. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98, pp. 275–281. Melbourne, Australia (1998)
42. Prévot, L., Borgo, S., Oltramari, A.: Interfacing ontologies and lexical resources. In: C. ren Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, L. Prévot (eds.) Ontology and the Lexicon, a Natural Language Processing Perspective, Studies in Natural Language Processing, pp. 185,200. Cambridge University Press (2010)
43. Pylkas, K., Erkko, H., Nikkila, J., Solyom, S., Winqvist, R.: Analysis of large deletions in BRCA1, BRCA2 and PALB2 genes in Finnish breast and ovarian cancer families. BMC Cancer **8**, 146 (2008)
44. Ranwez, S., Ranwez, V., Villerd, J., Crampes, M.: Ontological distance measures for information visualisation on conceptual maps. In: R. Meersman, Z. Tari, P. Herrero (eds.) On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, *LNCS*, vol. 4278, pp. 1050–1061. Springer (2006)
45. Resnik, P.: 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)
46. Reynar, J.C.: Topic segmentation: Algorithms and applications. PhD thesis (2000)
47. Riedhammer, K., Favre, B., Hakkani-Tür, D.: Long story short ? Global unsupervised models for keyphrase based meeting summarization. Speech Communication **52**(10), 801–815 (2010)
48. Salton, G., Singhal, A., Buckley, C., Mitra, M.: Automatic text decomposition using text segments and text themes. Hypertext'96 pp. 53–65 (1996)
49. Schmid, H.: Treetagger. In TC project at the institute for Computational Linguistics of the University of Stuttgart (1994)
50. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in WordNet. In: R.L. de Mántaras, L. Saitta (eds.) ECAI, pp. 1089–1090. IOS Press (2004)
51. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 888–905 (1997)
52. Staab, S., Maedche, A.: Ontology learning for the semantic web. IEEE Intelligent Systems **16**(2), 72–79 (2001)
53. Stokoe, C., Oakes, M.P., Tait, J.: Word sense disambiguation in information retrieval revisited. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03, p. 159. Toronto, Canada (2003)
54. Supekar, K., Chute, C.G., Solbrig, H.: Representing lexical components of medical terminologies in OWL **2005**, 719–723 (2005)
55. Sy, M., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., Ranwez, V.: User centered and ontology based information retrieval system for life sciences. BMC Bioinformatics **13**(Suppl 1), S4 (2011)
56. Wiss, U., Carr, D.: A cognitive classification framework for 3-Dimensional information visualization (1998)
57. Xie, S., Hakkani-tür, D., Favre, B., Liu, Y.: Integrating prosodic features in extractive meeting summarization. In: In Proceedings IEEE Workshop on Speech Recognition and Understanding (ASRU) (2009)
58. Zheng, H., Borchert, C., Jiang, Y.: A knowledge-driven approach to biomedical document conceptualization. Artificial Intelligence in Medicine **49**(2), 67–78 (2010)