Exploiting Semantics for Searching Agricultural Bibliographic Data

Journal of Information Science
1–15
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Domenico Beneventano
Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Italy

Sonia Bergamaschi
Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Italy

Riccardo Martoglia
FIM Department, University of Modena and Reggio Emilia, Italy

Abstract
Filtering and search mechanisms which permit to identify key bibliographic references are fundamental for researchers. In this paper we propose a fully automatic and semantic method for filtering/searching bibliographic data, which allows users to look for information by specifying simple keyword queries or document queries, i.e. by simply submitting existing documents to the system. The limitations of standard techniques, based on either syntactical text search and on manually assigned descriptors, are overcome by considering the semantics intrinsically associated to the document/query terms; to this aim, we exploit different kinds of external knowledge sources (both general and specific domain dictionaries or thesauri). The proposed techniques have been developed and successfully tested for agricultural bibliographic data, which plays a central role to enable researchers and policy makers to retrieve related agricultural and scientific information by using the AGROVOC thesaurus.

Keywords
Bibliographic Data; Agricultural Thesaurus; Semantic Knowledge Management; Semantic Similarity

1. Introduction

Agricultural bibliographic data plays a central role to enable researchers and policy makers retrieve related agricultural and scientific information and data [1]. The AGRIS[1] (International System for Agricultural Science and Technology) system, which currently contains more than 7 million bibliographic references on agriculture research and technology, seeks to use bibliographic data as an aggregator of locating not only the full-text of the article, but also related content across information systems available on the Web.

Moreover, there is a high number of databases publicly available for the agricultural research community and one of the most relevant information available in such systems is the bibliography associated to the instances of the databases. Filtering and search mechanisms which permit to identify key bibliographic references, are fundamental for researchers. While advanced capabilities are available in such systems to search for database objects, search mechanisms that allow querying bibliographic references are generally limited to free-text search and manually assigned descriptors.

We tested the relevance of performing bibliographic search within the CEREALAB[2] project, where we designed and developed a database containing genotypic and phenotypic data [2]; it helps cereal breeders for marker-assisted selection, e.g. for choosing molecular markers associated to economically important phenotypic traits. For example, the *Querying Wheat Genes* functionality, available in CEREALAB, allows users to find the wheat genes underlying a selected trait; the user can choose the trait 'Frost tolerance' to search for wheat genes that underlie the tolerance to frost, by obtaining 3 genes with the following related references (title):

Corresponding author:
Riccardo Martoglia, *FIM - University of Modena and Reggio Emilia, via Campi 213, 41125, Modena, Italy*
Riccardo.martoglia@unimo.it

(1) [BD1] Location of a gene for frost resistance on chromosome 5A of wheat;
(2) [BD2] The cold-regulated transcriptional activator Cbf3 is linked to the frost-tolerance locus Fr-A2 on wheat chromosome 5A;
(3) [BD3] Mapping genes affecting flowering time and frost resistance on chromosome 5B of wheat.

Let us suppose now that a user want to search bibliography related to *"Wheat"* and *"Frost tolerance"* on other databases that as CEREALAB contain genotypic and/or phenotypic data, such as GrainGenes [3] and Gramene [4]. In such systems, in order to manage and retrieve bibliographic references, mechanisms based on either text search techniques and on manually assigned descriptors are generally adopted. Standard syntactical search techniques [5] (roughly speaking, these techniques look for documents containing and/or annotated with the same terms specified by the user query) often suffer of low effectiveness as they are inadequate to capture the similarity between documents and disregard the semantic connections (*synonyms or semantic relations*) of the terms composing them. For instance, in a syntactical text search approach, no match would be found between the title of [BD2] and the other titles [BD1] and [BD3], as they have no words in common.

Similar problems are found with search mechanisms based on manually assigned descriptors. For instance, in GrainGenes [3], a database similar to CEREALAB, individual bibliographic references are manually indexed by using a set of keywords; then a query interface[3] is provided to search for references; as an example, by specifying the keyword query *"Wheat"* and *"Frost tolerance"*, i.e., searching for reference records associated to both these two keywords, we obtain 7 records, among which there are [BD2] and [BD3] but not [BD1]. On the contrary, by adding semantics, i.e., synonyms and related terms (i.e., broader, narrower or correlated terms), it is possible to discover that in AGROVOC[4] thesaurus the term "Frost tolerance" is a synonym of "Frost resistance" and, thus, [BD1] and [BD3] can be related to the keyword query *"Wheat"* and *"Frost tolerance"*.

Let us suppose now that a user want to search bibliography related to *"Wheat"* and *"Frost tolerance"* by using the AGRIS system, that is among the most comprehensive online collections of agricultural and related sciences information. Bibliographic records in AGRIS are indexed by AGROVOC terms and then such terms are used to formulate search queries; for example, *"Wheat"* and *"Frost tolerance"* are two AGROVOC terms and we may use them to formulate a search query in AGRIS (i.e., by specifying they as *AND terms* in the query form available at http://agris.fao.org/agris-search/biblio.do). Such a query returns no record; on the other hand, by using *"Frost resistance"* instead of *"Frost tolerance"* (i.e. the keyword query with *"Wheat"* and *"Frost resistance"* as AGROVOC terms) we obtain 3 records but not [BD1] ([BD2] and [BD3] are not in the AGRIS collection), although its title contains the keyword *"Frost resistance"* and it is indexed with the AGROVOC term *"Temperature resistance"* which is a related term (*RT*) of *"Frost resistance"*. In other words, bibliographic records are indexed by AGROVOC terms but semantic relationships between terms are not fully exploited in search queries.

Starting from these considerations, we propose a fully automatic and semantic method for filtering/searching bibliographic references, which allows users to look for information by specifying simple keyword queries or document queries (i.e. by simply submitting existing documents to the system). The proposed approach *effectively* identifies the similarities between keyword queries and a reference set of documents: the limitations of standard syntactical techniques are overcome by considering the *semantics* intrinsically associated to the document/query terms; to this aim, we exploit different kinds of external knowledge sources (both general and specific domain dictionaries or thesauri). Moreover, the proposed solution does not require big investments or knowledge prerequisites: first of all, it exploits the large amounts of bibliographic documents already available in each research group or bibliographic collection, without requiring any conversion towards complex structured formats which would be time and cost consuming. Secondly, manual annotation of the processed papers is also not required, since the method works on the main textual contents of the papers; at the same time, it can also be easily applied to already annotated collections, leveraging the work of the experts in the field. In either case, thanks to the automatically identified semantic relationships, the high effectiveness of the search process is ensured.

The techniques are an improved version of the preliminary ones we originally employed in a software development context ([6]). In this paper, we extend and customize them to the plant science domain by considering the exploitation of the AGROVOC thesaurus potentialities, a new term similarity formula and we also design a complete relational data structure supporting their execution in an effective and efficient way.

Although the methodology has been applied for filtering/searching bibliography documents, it is general and can be applied to other document collections of the plant science and agricultural domain. On the other hand, as discussed in [7], agricultural bibliographic data plays a central role to enable researchers and policy makers retrieve related agricultural and scientific information and data.

The paper is organized as follows. Section 2 discusses related work. Section 3 introduces the **Semantic Engine** component, which implements the proposed method. In Section 4, we focus on the analysis and semantic techniques on which the **Semantic Engine** component is based. The detailed experimental evaluation presented in Section 5 shows the achieved effectiveness results, going beyond typical retrieval solutions. Finally, Section 6 concludes the work and gives an outlook on future work.

2. Related Work

A broad range of methods for semantic document retrieval has been developed in the context of the Semantic Web, as discussed in [8], a survey which covers approaches that exploit domain knowledge to process search requests; the authors present a large variety of domain knowledge utilisations that comprises automatic query expansion and ontology-driven document retrieval.

The relative ineffectiveness of information retrieval systems is largely caused by the inaccuracy with which a query formed by a few keywords models the actual user information need; one well known method to overcome this limitation is automatic query expansion, whereby the user's original query is augmented by new features with a similar meaning [9]. One of the most natural approaches to automatic query expansion is the manipulation of textual data, such as using a stemming algorithm to reduce different words to the same stem and finding synonyms of a query word from a thesaurus, most usually from WordNet [10]. More specifically, similar semantic techniques have also been exploited in a bibliographic search context and for particular domains/objectives; in [11] automatic query expansion techniques are evaluated in the context of bibliographic data searching, with reference to MEDLINE, a well-known premier bibliographic collection that contains references to articles contained in journals on life sciences. Both in general and bibliographic contexts, and differently from our approach, complex query expansion techniques such as the ones discussed usually require different parameters to be specified (as also stated in [11]). For example, in [10] the parameter set for a given run specifies for each relation type included in WordNet the maximum length of a chain of that type of link that may be followed. Generally, there is no single theory capable of finding the most appropriate values [11] and therefore a long process of manual "tuning" becomes necessary.

An increasing number of document retrieval systems make use of ontologies to help users clarify their information needs and come up with semantic representations of documents. In order to simplify cross-referencing between search results, in [12] an ontology-augmented bibliographic search engine was developed where the regular, keyword based bibliographic search is improved with the help of an ontology which facilitates the resolution of ambiguities among query keywords. In a strictly medical domain, and with the aim of achieving graphical result browsing, [13] investigates how keyword search can be enhanced through the use of the Gene Ontology, a hierarchically structured vocabulary for molecular biology, to structure the large amounts of biomedical literature contained into the PubMed[6] database; the main contribution is the introduction of ontology-based literature search by creating GoPubMed, which submits keywords to PubMed, extracts Gene Ontology-terms from the retrieved abstracts, and presents the relevant sub-ontology for browsing. In [14], ontologies are exploited for query formulation, query routing and answer presentation into Bibster, a Peer-to-Peer system for exchanging bibliographic data among researchers.

As to the employed knowledge sources, most of the works focus on a single source (for instance, WordNet [10] or Gene Ontology [13]). In [15], a Simple Knowledge Organization System (SKOS) based term expansion and scoring technique that leverages labels and semantic relationships of SKOS concept definitions is proposed. In this case, the expansion technique is configurable w.r.t different used vocabularies, however one dictionary at a time is employed. Instead, in our approach, we concurrently exploit multiple knowledge sources, i.e. WordNet and AGROVOC, in order to provide a satisfying coverage of different categories of terms.

Moreover, focusing on the necessity of manual intervention, typical semantic retrieval techniques obtain good effectiveness levels only on manually annotated collections and/or with explicit user intervention. An evaluation of bibliographic database access using free-text and controlled vocabulary is in [16], which evaluates the retrieval effectiveness of various search models, based on either automatic syntactic text-word indexing or on manually assigned controlled descriptors. In [17], a query expansion technique works on MEDLINE documents which have been manually assigned to controlled MeSH (Medical Subject Headings) vocabularies. The advanced indexing and retrieval method we proposed in this paper, instead, exploits the semantics of the text while remaining completely automatic.

In order to conclude our analysis, we will now consider the specific context of databases publicly available for the agricultural research community. Also in this case, filtering and search mechanisms that allow querying bibliographic references are generally limited to free-text search and manually assigned (see, for instance, GrainGenes [3] and

Gramene [4]). In such systems, in order to manage and retrieve bibliographic references, mechanisms based on either text search techniques and on manually assigned descriptors are generally adopted.

Some agricultural databases make also available advanced semantic capabilities to search for database instances. As an instance, for the design and the development of the CEREALAB Database, we employed semantic methods and technologies offered by the MOMIS Data Integration system [18]. In [2] we presented a methodology to publish and link the CEREALAB database to the Linked Open Data cloud, in order to facilitate breeders and geneticists in searching and exploiting linked agricultural resources. In [19] the author, starting from the fact that the breadth of biodiversity literature available through the Biodiversity Heritage Library (BHL) is potentially of great use to agricultural research, explores the practical issues arising from attempting to filter out relevant legacy literature to support agricultural research; they conclude that the breadth of coverage of BHL can complicate finding relevant literature and then highlight the importance of using metadata and semantic search.

In [20] a retrieval engine which exploits the content and structure of available domain ontologies to expand and enrich retrieval results in major plant genomic databases is proposed. We agree with the author's statement that, though much time and effort have been spent on the development of plant-related ontologies, the knowledge embedded in these ontologies remains largely unused in available plant search mechanisms. In the context of agricultural bibliographic data, one exception is the AGRIS system [1], where bibliographic records are indexed by AGROVOC terms and the end-user can use such terms to formulate search queries; however, as discussed in the introduction, search queries on the AGRIS system rely heavily on manual indexing, i.e., selected terms of the AGROVOC thesaurus are assigned to each publication by a human indexer.

Generally speaking, our discussion once again confirms that most of the drawbacks we discussed for semantic document retrieval and bibliography querying systems are also present in systems specifically working in an agricultural data scenario. On the other hand, the method we proposed uses the AGROVOC and WordNet thesauri in a completely different way: first of all, no human intervention is required (either in document indexing or retrieval); the search techniques fully exploit the semantic relations available in the thesauri; moreover, the retrieval and ranking model is seamlessly extended in order to manage semantic features and/or advanced functionalities (such as automatic composite terms identification, which is also absent in most of the discussed approaches).

3. Semantic Document Analysis and Suggestion Search Techniques

The method has been implemented in the **Semantic Engine** component. The goal of the Semantic Engine is to analyze and search the relevant textual information available in a document collection. The input of the search mechanism is the query submitted by the users. Two main usage scenarios are encompassed:

- **a "from topics" scenario**, where users submit a short text or keyword list describing the topics and contents they are interested in;
- **a "from example" scenario**, where users provide an existing document (e.g. an article selected from the collection, or even a whole new document) which should be used as an "example" to find documents with similar topics/contents.

Then, through a suggestion search phase, the user receives a set of suggestions in the form of the most relevant documents available in the collection. To this end, the Semantic Engine supports two main processes (see Figure 1):

(1) **Document collection Analysis and Semantic Index Population:** during this offline process, the Semantic Engine automatically extracts and "normalizes" the informative contents of the given set of documents (i.e., the articles in a given collection) in the form of a shared "terminology", eventually populating the computer-processable Semantic Index; such information will be used during the online (document retrieval) process. In particular, the extracted terminology is also enriched with statistical and semantic information (i.e., links to thesauri and domain vocabularies, definitions, and synonyms).

(2) **Relevant Document Suggestion Search:** in this online process, user queries are processed and relevant documents are identified. First of all, the query text is analyzed by means of the same techniques used for the Semantic Index population. Once the query has been reduced to a set of terms with associated semantics, appropriate semantic similarity techniques are exploited to easily identify relevant documents in the collection, and to produce a list of suggestions ranked on the similarity (relevance) score.

The semantic text analysis phase, involved in both of the processes described above, the employed knowledge sources and the structure of the semantic index produced in the offline process are detailed in Section 4.1; the semantic similarity computation phase, involved in the online process, is detailed in Section 4.2; the data structure of the semantic index is in Appendix.

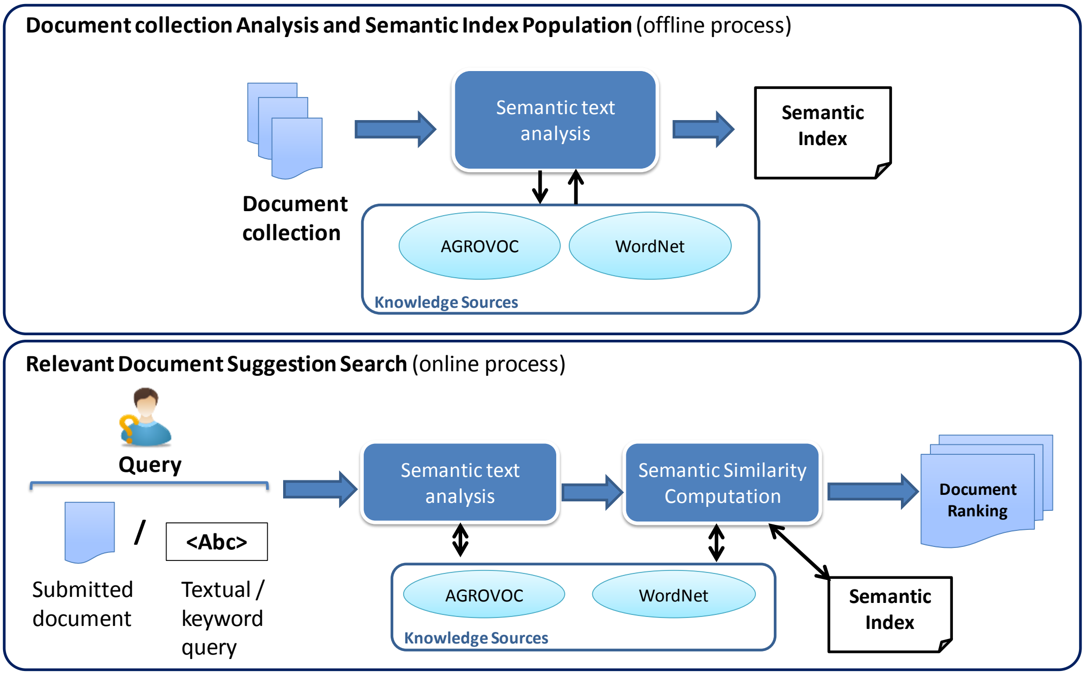


Figure 1. Semantic Engine offline (top part) and online (bottom part) processes.

4. Semantic Engine Techniques

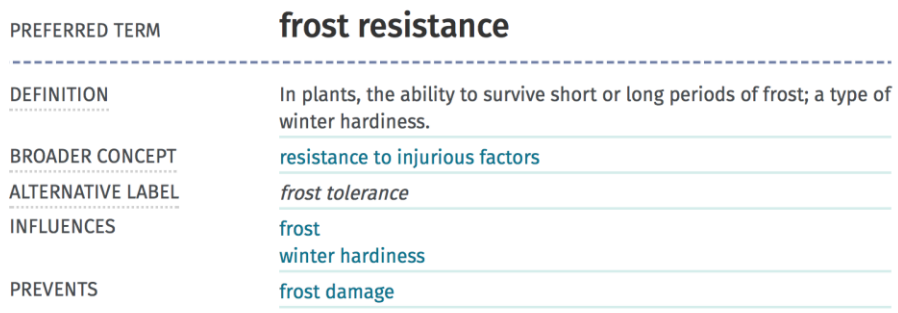
4.1. Semantic Text Analysis and Employed Knowledge Sources

Our goal for semantic text analysis was to devise a flexible technique to be exploited both for “off-line" analysis (thus working on the documents already available in document collection) and for “on-line" querying operations, i.e., applied on the fly to the submitted queries. The semantic text analysis phase involves some preliminary pre-processing steps, i.e. *tokenization*, where terms are identified and punctuation is removed, and *stemming*, where the tokens are “normalized” and “stemmed”, i.e., terms are reduced to their base form (managing plurals and inflections). Then, we perform *POS (Part of Speech) Tagging*, i.e. the tokens are “tagged” with Part of Speech tags (noun, verb, ...). The latter allows the semantic engine to perform the next step, which can be key to the effectiveness of the subsequent search process: *composite terms identification*, where possible composite terms (such as “farm forestry” or “frost tolerance”) are identified by means of a simple state machine and of the computed POS tags information.

After those preliminary steps, the Semantic Engine exploits the power of external knowledge sources in a *filtering and enrichment* step: the most relevant terms are selected and associated to additional semantic information:

(1) **Synonyms**: equivalent terms that can be interchanged in our agricultural context, e.g. “Frost damage” and “Frost injury”;

(2) **Related Terms**: terms that are not equivalent but nonetheless can be very relevant with respect to a term. These include: *broader terms* and *narrower terms* (e.g. “paramecium” belongs to “protozoa”) or *correlated terms* (e.g. “winter hardiness” relates to “frost tolerance”).



PREFERRED TERM	frost resistance
DEFINITION	In plants, the ability to survive short or long periods of frost; a type of winter hardiness.
BROADER CONCEPT	resistance to injurious factors
ALTERNATIVE LABEL	*frost tolerance*
INFLUENCES	frost winter hardiness
PREVENTS	frost damage

Figure 2. An excerpt of the AGROVOC Thesaurus for the term "Frost Resistance"

More specifically, we made use of the AGROVOC Thesaurus, covering specialist terms in the agricultural area, and the WordNet English thesaurus [21], complementing the specialistic source with general knowledge about English concepts. The coverage of the two sources is quite diverse. AGROVOC mainly includes specialistic terms belonging to the agricultural domain, while WordNet covers a wider range of domains. While WordNet misses several specific terms belonging to the considered domain (e.g., "crop production"), it can definitely help by bringing additional knowledge on more common concepts that can be present in documents (e.g., "income", "documentation", etc.). For instance, Figure 2 shows an excerpt of the AGROVOC Thesaurus for the term "Frost Resistance": we can easily derive that it is related to the broader concept of "Resistance to Injurious Factors" and to other correlated terms such as "Frost" and "Winter Damage"; moreover the synonym "Frost Resistance" (denoted by Alternative Label) is given.

The information composing our semantic index are completed by a final *term statistics and weight computation* step, where specific weights are computed for each term, reflecting their relevance and meaningfulness in the document. As in classic Information Retrieval, besides term frequency (TF), we compute the inverse document frequency (IDF) [22][5], which provides an estimate of the meaningfulness of each term. The weight is then computed as TF*IDF. In this way, common terms, which are present in a large number of documents, have a lower weight and will give a lower contribution to the final similarity, since they are probably less meaningful in the context we consider. By means of the weights, the content of the index allows the similarity functions of the Semantic Engine to draw useful knowledge from both the semantic and the text retrieval research areas.

Table 1. A sample portion of the extracted Semantic Index (global view).

TERM	AGR	WN	SYNONYMS	RELATED	IDF	DOC_LIST
Farm forestry	Y	N	Agroforestry, …	*Broader:* Farming systems, …	7.4570	['D02865', 'D06531', …]
Frost tolerance	Y	N	Frost resistance, …	*Broader*: Resistance to injurious factors, … *Correlated*: Frost, Winter hardiness, …	4.5643	['D01356', …]
Income	Y	Y	Profit, …	*Narrower*: Farm income, … *Correlated*: Cash flow, …	3.5835	['D00342', 'D00789', …]

By applying batch semantic text analysis to the document collection, the **Semantic Index** is automatically generated. Conceptually, the Semantic Index consists of a **global view** (all terms in all documents, together with their statistics, see Table 1 for a small example) and a **per-document view** (terms occurrences in each document with their statistics, Table 2). In particular, **DOC_LIST** is the list of the documents IDs in which each term occurs.

Table 2. A sample portion of the extracted Semantic Index (per-doc view).

DOC	TERM	TF	WEIGHT(TF*IDF)
D00001	Flowering	0.5455	0.9773
D00002	Enzyme activity	0.2105	1.1316
D00002	Fruit	0.0526	0.1262

4.2. Semantic Similarity Computation

Since the need of effectively and efficiently computing similarities between documents is crucial in our context, we want to exploit a *document similarity formula* $DSIM(D^x, D^y)$ which, given a source document $D^x = \{t_i^x, \dots, t_n^x\}$ and a target document $D^y = \{t_i^y, \dots, t_m^y\}$, quantifies the similarity of the source document w.r.t. the target document. Typically, the source document will be represented by the query information submitted by the user, while target documents will be the ones available in the document collection. Being documents represented by sets of terms, semantic similarity computation becomes a matter of computing similarities between sets of terms. Therefore, we exploit a variation of the similarity framework originally proposed in [6], where document similarity, in turn, uses a combination of the scores provided by a *term similarity formula TSim* between the document terms:

$$DSIM(D^i, D^j) = \sum_{t_i^x \in D^x} TSim\left(t_i^x, t_{\bar{j}(i)}^y\right) * w_i^x * w_{\bar{j}(i)}^y \tag{1}$$

where $w_i^x = tf_i^x * idf_i$ and $w_{\bar{j}(i)}^y = tf_{\bar{j}(i)}^y * idf_{\bar{j}(i)}$ and

$$t_{\bar{j}(i)}^y = argmax_{t_j^y \in D^y}\left(TSim\left(t_i^x, t_j^y\right)\right) \tag{2}$$

Each term contributes to the final similarity with a different weight, i.e., more frequent and more significant terms contribute more to the similarity between the two documents. In our agricultural context, *TSim* is computed by means of Equation (3) which considers the semantic information available in the Semantic Index, i.e. synonyms and related terms extracted from AGROVOC and/or WordNet:

$$TSim\left(t_i, t_j\right) = \begin{cases} 1, \ if \ t_i = t_j \ or \ t_i \ SYN \ t_j \\ r_1, \ if \ t_i \ REL_{BR} t_j \\ r_2, \ if \ t_i \ REL_{COR} \ t_j \\ 0, \ otherwise \end{cases} \tag{3}$$

More specifically, the case of maximum similarity (value 1) corresponds to the case where the two terms are synonyms (*SYN* relation). Moreover, the formula provides further cases where the two terms are not equal or synonyms, nonetheless they are in some way related from a semantic point of view: REL_{BR} for broader terms and REL_{COR} for correlated terms. Such terms will contribute with a user-defined fixed similarity value r_i (i=1…2, 0<r_i<1). Note that, by means of REL_{BR}, a document term t_j (e.g. “forest fire management”) is considered similar to broader query terms t_i (e.g. “forest protection”) but not to narrower ones, since a user interested in a general topic could be most likely interested in its more specific instantiations (but not vice-versa).

In this way, a *ranking* of the available documents (on the basis of *DSim*) is induced, thus predicting which documents are relevant and which are not w.r.t. D^x.

Let us now see how the Semantic Engine techniques we just presented can be used in the context of a small illustrative example. Let us suppose that *D*1 is a fragment of a document available in the collection:

*D*1. “Novel **fodder mechanization** systems for Hawaii"

Given the following query:

$Q1$. “**Modernization** of **forage** cultivation and processing techniques"

With a syntactic search approach, no match should be found between the query and the document: $D1$ has no terms in common with $Q1$. However, by adding semantics and using the Semantic Engine techniques, we can easily determine that $D1$ might contain information potentially relevant to $Q1$, as “fodder" is a synonym of “forage", those terms will thus likely give a strong contribute to the similarity between $Q1$ and $D1$. Moreover, the term “mechanization" can be found as related to “modernization" (in fact, it is a narrower term, i.e. mechanization is a kind of modernization), thus this match will also contribute to the relevance of $D1$, even if less significantly (related terms usually contribute to weaker scores than synonyms or equal terms).

5. Experimental Evaluation

We will now present the results of the effectiveness evaluation we performed on the proposed method. We formed a representative collection of approximately 1800 textual documents, i.e. abstracts of published scientific papers in the agricultural domain. More specifically, we extracted them by extracting from the CEREALAB bibliography documents on ten topics which are common in the domain, such as “land preparation”, “mixed cropping” and “plant habit”, and by selecting the first 200 results for each of them. Starting from this collection, we automatically generated the Semantic Index and we considered, with the help of experts in the domain, a large number of queries (nearly 100) simulating possible “from topics" scenario requests. Among them, we selected a set of 10 queries ($Q1-Q10$) as the most representative ones. Table 3 shows the complete specification and the main keywords contained in $Q1-Q10$ queries. Each query will be submitted to the current implementation of the semantic engine, so to generate a set of possible “suggestions”, i.e. pointers to the relevant documents in the collection. In order to evaluate the effectiveness of our approach, for each query the output of the engine will be compared to a “gold standard”, i.e. the relevant answers which were manually selected from the collection by experts of the agricultural domain. Moreover, we also considered additional queries composed by whole documents, i.e. “examples” selected by users that should be used to find documents with similar topics/contents, as in the “from example” scenario. The results obtained on a selection of these queries ($QT1-QT4$) will also be later analyzed in this evaluation.

Table 3. The keyword queries we selected for the Semantic Engine evaluation.

QUERY	QUERY SPECIFICATION	MAIN KEYWORDS EXTRACTED FROM QUERIES
Q1	“ provincial and national gathering of information on farmlands ”	Farmland, gathering, …
Q2	“ fodder for animals in different landscape types ”	Fodder, landscape, …
Q3	“ value of agricultural machinery and increase in income ”	Income, machinery, …
Q4	“ local expenditure and effects on land required for food ”	Expenditure, land, …
Q5	“ loss of habitats and species diversity and trust on animal power ”	Animal power, species diversity, …
Q6	“ farm forestry and practices for plant establishment in hill and upland regions ”	Farm forestry, plant establishment, …
Q7	“ risk analysis on alkaloids by mass spectrometry in food processing ”	Alkaloids, risk analysis, …
Q8	“ studies on effects of water availability on crop performance ”	Water availability, crop performance, …
Q9	“ techniques for crop production and pest management in general plant ecology ”	Crop production, ecology, …
Q10	“ substance adsorption in soil receiving fertilizer application and impact on livestock”	Fertilizer application, livestock, …

First of all, we will assess precision, recall, and F-measure of our topic queries: the left part of Table 4 presents a summarization of the results. In particular, for ease of analysis, Table 4 presents the results of each query Q1-Q10, along with their collective average and standard deviation figures; moreover, for completeness sake, it also shows average and standard deviation for all the remaining queries we considered in our study (Q11-Q95), therefore confirming the results we will discuss in detail for Q1-Q10. In order to emphasize the contribution of the different applied techniques to the achieved results, in the right part of table we also present the results concerning two important baselines: a standard “syntactic” retrieval method ignoring semantic synonyms and related terms, representative of document retrieval techniques commonly exploited by agricultural bibliographic research tools and systems, and another method not exploiting the text analysis phase (including stemming and composite terms identification). As we can see, the precision and recall levels achieved by the semantic method are generally very satisfying: all queries greatly benefit from semantic features such as synonyms and related terms management. See also Figures 3, 4 and 5 for a graphical precision, recall, and F-measure result comparison, respectively. Let us now analyze the results in detail for each of the queries Q1-Q10.

Table 4. Effectiveness analysis: precision, recall and F-measure (semantic engine results on the left, two baselines on the right).

	OUR RESULTS			TYPICAL RETRIEVAL BASELINES					
QUERY				NO SEM SYN/REL			NO KW SEL		
	PREC	REC	F	PREC	REC	F	PREC	REC	F
Q1	1.000	1.000	1.000	1.000	0.727	0.842	1.000	0.724	0.840
Q2	1.000	1.000	1.000	1.000	0.288	0.448	1.000	0.288	0.448
Q3	1.000	1.000	1.000	1.000	0.797	0.887	1.000	0.745	0.854
Q4	1.000	1.000	1.000	1.000	0.097	0.176	1.000	0.097	0.176
Q5	1.000	1.000	1.000	1.000	0.714	0.833	0.167	0.768	0.274
Q6	1.000	1.000	1.000	1.000	0.091	0.167	0.021	0.134	0.036
Q7	1.000	1.000	1.000	1.000	0.200	0.333	1.000	0.200	0.333
Q8	1.000	1.000	1.000	1.000	0.143	0.250	0.029	0.973	0.057
Q9	0.965	1.000	0.982	1.000	0.101	0.184	0.605	0.854	0.708
Q10	0.958	1.000	0.979	1.000	0.645	0.784	0.122	0.673	0.206
AVG	0.992	1.000	0.996	1.000	0.380	0.490	0.594	0.546	0.393
STDEV	0.015	0.000	0.008	0.000	0.285	0.294	0.433	0.312	0.292
(Q1-Q10)									
AVG	0.962	1.000	0.980	1.000	0.354	0.523	0.522	0.553	0.412
STDEV	0.021	0.000	0.013	0.000	0.299	0.301	0.468	0.311	0.299
(Q11-Q95)									

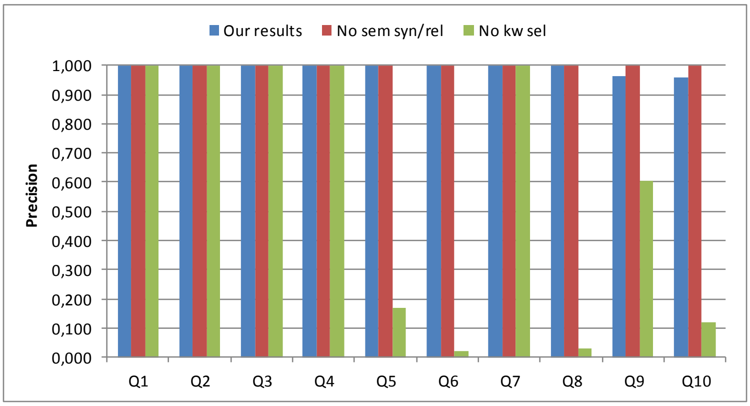


Figure 3. Detailed precision graph for queries Q1-Q10

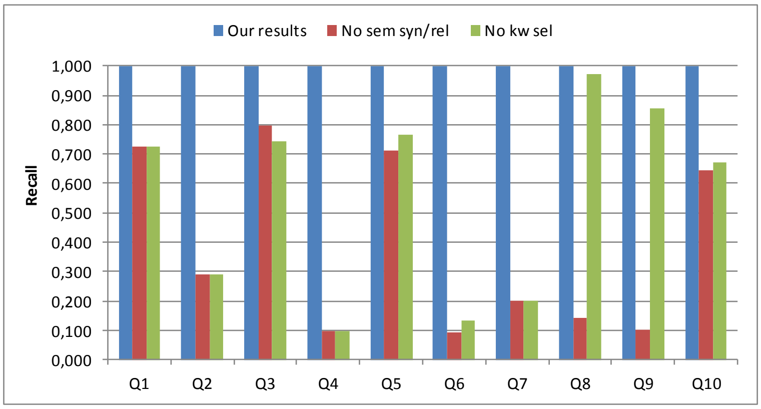


Figure 4. Detailed recall graph for queries Q1-Q10.

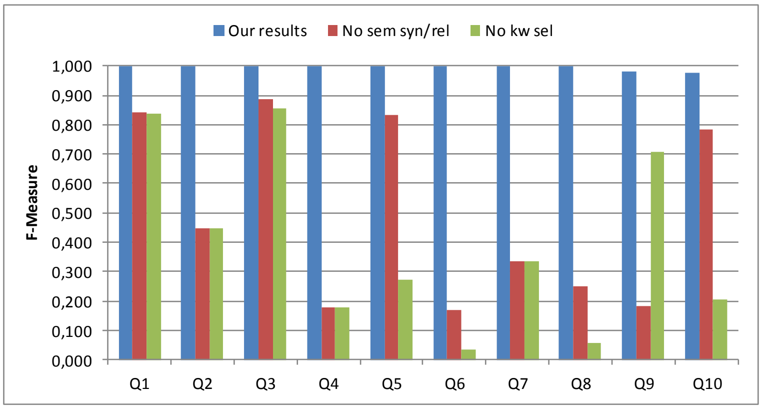


Figure 5. Detailed F-measure graph for queries Q1-Q10

The first queries (*Q*1 to *Q*6) mainly benefit from our synonym management techniques. For instance, *Q*1 contains the term “farmland”: this form is generally one of the most commonly used for this concept, and this is also proven by the decent level of recall offered by the syntactic baseline. Anyway, by exploiting AGROVOC and the semantic engine, we know that “cropland” is a synonym and that also documents containing this term are relevant: in this case several more useful documents can be retrieved and presented to the user (see the recall figures in the left part of table). In *Q*2, the effect of synonym management is even stronger: for instance, the query term “fodder” is less common than its AGROVOC synonym “forage”, therefore in this case the amount of relevant documents that would be left out of the results of a syntactic search (thus, including all typical bibliographic search tools) would be much higher (see recall of 28% in the first baseline).

In some cases, queries contain some terms that are not necessarily specialized in the agricultural domain, but whose alternative forms can nonetheless be useful in order to make the retrieved results more complete: this is the case, for instance, of queries *Q*3 and *Q*4, containing general terms such as “income” and “expenditure”. By accessing our second knowledge source, WordNet, additional useful synonyms such as “profit” and “consumption” can be exploited in the retrieval process, allowing significantly higher recall levels than in simple syntactic search.

Queries *Q*5-*Q*10 require a wide range of processing techniques in order to be fully satisfied: this is why the gap between the results of the semantic engine and those of the other baselines gets even wider. For instance, automatically and correctly identifying composite terms (such as “animal power” in *Q*5 and “farm forestry” in *Q*6) can be key to the precision of the retrieved results: this is evident from the very low precision results that are achieved by the second baseline for *Q*5 and *Q*6, where text analysis (including composite term identification) is not performed (16% and 2%, respectively). This is due to the fact that many irrelevant documents (for instance, those containing “animal” and "power", but not used together as a composite term) are retrieved and presented. Moreover, by correctly identifying composite terms, the semantic engine can also exploit the power of synonyms and related terms on them (e.g., "agroforestry" for *Q*6).

Queries from *Q*7 to *Q*10 are quite specific and their processing can significantly benefit also from the semantic engine related terms. For instance, "water scarcity" is strongly correlated to "water availability": not retrieving documents involved with "water scarcity" has a very negative impact on recall (14% in the syntactic baseline). On the other hand, simply retrieving all documents containing "water" (second baseline) certainly produces good recall but terrible precision figures, due to the very common nature of the word considered alone. A very important role is also played by broader and narrower terms, as extracted from AGROVOC: a query such as *Q*7 involves "alkaloids", thus documents related to specific alkaloids, such as "anabasine" or "colchicine", are certainly relevant but not always easy to discover without semantic technology. The same holds, for instance, for *Q*9, where by specifying "crop production" users are typically interested in the many included specific activities (e.g. "cultivation", "drilling", "harvesting", etc.). In these cases the recall offered by syntactic search is very unsatisfying (20% or lower), since many of the documents are not retrieved. Please note that expanding search to related terms could potentially increase the probability of retrieving irrelevant documents; anyway, in our agricultural setting, the precision offered by the semantic engine and by its use of the AGROVOC knowledge source is kept very high (97% and higher).

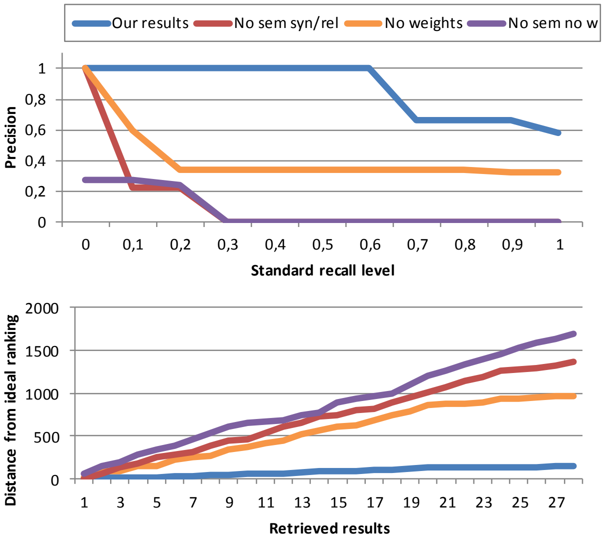


Figure 6. In-depth effectiveness analysis for QT1: precision at standard recall levels (top) and distance from optimal ranking (bottom).

Finally, we deepened the effectiveness analysis by considering actual text documents (*QT*1-*QT*4) for which to find related documents in the collection, in "from example" scenarios. Such queries contain a very large number of terms and can possibly produce a lot of results: thus, it is essential to evaluate also the induced ranking, so to assess whether the best suggestions are returned in the top positions and, thus, whether the proposed weighting scheme is effective. Figures 6 and 7 show (upper part) the precision values obtained for *QT*1 and *QT*2 (other queries performed very similarly) at different recall levels, i.e., when a given percentage of relevant documents have been found, and the distance from the ideal ranking (lower part). The results of the semantic engine are compared to the syntactic baseline not considering synonyms and related terms, a non-weighted approach, and a non-weighted syntactic approach. Notice that, in both cases, our technique achieves very high precision levels even at high recall levels: for instance, at recall level 0.6, the precision is still 1 and 0.8 for *QT*1 and *QT*2, respectively, while the baselines precision levels have already dropped lower than 0.4 and 0.2. This confirms that our techniques are able to identify the most significant terms in the "example" documents, without being misled by non-relevant ones. The optimal ranking distance analysis confirms the goodness of the retrieved results: the curve represents the normalized Spearman footrule distance [23] between the retrieved and the ideal ranking, i.e. the normalized sum of the absolute values of the difference between the ranks. In this case, the curves of the semantic engine are the lowest ones, meaning the least distance to the optimal ranking.

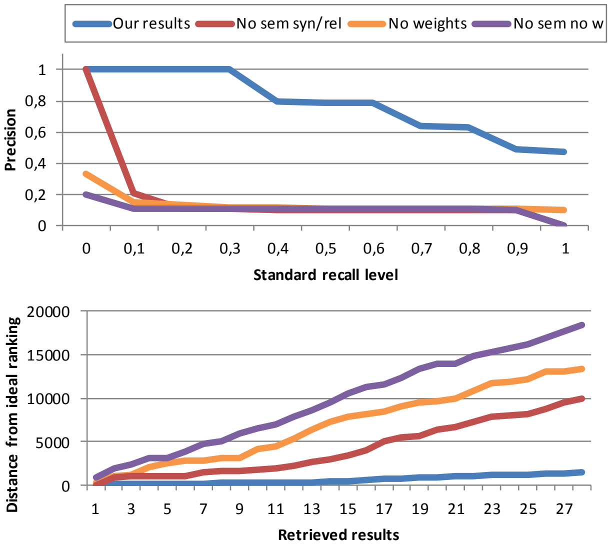


Figure 7. In-depth effectiveness analysis for QT2: precision at standard recall levels (top) and distance from optimal ranking (bottom).

In conclusion, we can observe that, in all cases, our semantic method leads to improvements in precision without compromising recall and without requiring any manual effort (e.g. manual annotation of the documents). Moreover, it gives also a key contribution in retrieving the most relevant results as first in the ranking.

6. Conclusion and Future Work

In this paper we proposed a fully automatic and semantic approach for filtering/searching bibliographic data, carefully considering the agricultural research community. The main achievements of the proposed approach, as also documented by the experimental section, are the following:

- The tool leverages on the strengths of both classic information retrieval and of knowledge-based techniques and is able to automatically identify the similarities between keyword queries and a collection of bibliographic references. The limitations of standard syntactical techniques are overcome by considering the *semantics* intrinsically associated to the document/query terms.
- The approach does not have any prerequisite, such as the knowledge of complex formal representation/querying standards, or the configuration of complex runtime parameters, and does not require any conversion towards complex structured formats, thus proving both time and cost effective.
- Differently from most of the available semantic retrieval techniques, manual annotation of the processed bibliographic data and/or user intervention is not required, since the method directly works on the main textual contents of the papers; at the same time, it can also be easily applied to already annotated collections, leveraging the work of the experts in the field.
- We exploit advanced text analysis and both general and specific domain dictionaries or thesauri, showing that, contrarily from many of the available bibliographic and semantic search techniques, the use of such techniques as automatic composite terms identification and the concurrent use of more than one knowledge source is feasible (and advisable) in a specific setting such as the considered agricultural one.

Several paths will be contemplated as future work:

- We will further analyze and refine the similarity techniques, user feedback on the retrieved suggestions, multi-language information management and querying support. In particular, Word Sense Induction, i.e. the automatic discovery of word senses from raw text, has been shown useful in many scenarios; in the context of an Intelligent Search scenario, Word Sense Induction, was proved as a novel approach to Web search result clustering [24]. We are currently developing an innovative Word Sense Induction method based on multilingual data; the method can be exploited for Multilingual Web Access concerned with retrieval from the Web, where documents in multiple languages co-exist and need to be retrieved to a query in any language. As future work, we will test such technique in the context of the agricultural domain, also based on the fact that the AGROVOC thesaurus is multi-lingual.
- We will extend our approach towards a distributed environment. In particular, the AGRIS network is currently deliberating on a proposal for a semantically rich OpenArchives architecture in the area of Agricultural Sciences and Technology; the implementation of the new architecture will lead to a network of hundreds of open archives [1]. Jointly exploiting this enlarged citation information following our semantic approach will possibly lead to further progresses in coverage and effectiveness.
- Lessons learned from the development of a semantic bibliographic search engine may facilitate general semantic search engines. Although the approach has been studied and evaluated in the context of the CEREALAB project and used specific knowledge sources, such as AGROVOC and WordNet, the methodology is general and could be exploited with other bibliographic datasets and thesauri. Moreover, the approach could be applied and extended to other collections of textual documents; one notable example is the FAO collection, an online document repository[7] that is large and well used (1M hits/month) and where documents are currently manually indexed with terms from AGROVOC.

Notes

1. http://agris.fao.org/agris-search/index.do
2. http://www.cerealab.org
3. http://wheat.pw.usda.gov/GG2/quickquery.shtml#references
4. http://aims.fao.org/standards/agrovoc/
5. IDF is obtained by dividing the total number of documents by the number of documents containing the term and then by computing the logarithm of that ratio.
6. http://www.ncbi.nlm.nih.gov/pubmed
7. http://www.fao.org/documents

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

[1] Anibaldi S, Jaques Y, Celli F, et al. Migrating bibliographic datasets to the semantic web: the AGRIS case. Semantic Web http://www.semantic-web-journal.net/content/migrating-bibliographic-datasets-semantic-web-agris-case-0 (accessed 30 March 2015)

[2] Beneventano, D., Bergamaschi, S., Sorrentino, S., Vincini, M., Benedetti, F.: Semantic annotation of the CEREALAB database by the agrovoc linked dataset. Ecological Informatics 26(2), 2015, pp. 119-126.

[3] Carollo V, Matthews DE, Lazo GR, et al. Graingenes 2.0. An improved resource for the small-grains community. Plant Physiology 139(2), 2005, pp. 643–651

[4] Liang C, Jaiswal P, Hebbard C, et al. Gramene: a growing plant comparative genomics resource. Nucleic Acids Research 36(suppl 1), 2008, pp. 947–953.

[5] Baeza-Yates RA and Ribeiro-Neto B. Modern Information Retrieval. Addison- Wesley Longman Publishing Co., 1999.

[6] Martoglia, R.: Facilitate IT-Providing SMEs in Software Development: a Semantic Helper for Filtering and Searching Knowledge. In: SEKE, Knowledge Systems Institute Graduate School, 2011, pp. 130–136.

[7] Malapela T, Celli F, Subirats I, et al.: The role of agris in providing global agricultural information to boost productivity and food security. Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 140 - Agricultural Libraries Special Interest Group. In: IFLA WLIC 2014, 16-22 August 2014.

[8] Mangold, C.: A survey and classification of semantic search approaches. International .Journal of Metadata, Semantic and Ontologies 2(1), 2007, pp. 23–34.

[9] Carpineto, C. and Romano, G.: A survey of automatic query expansion in information retrieval. ACM Computer Survey. 44(1), 2012, pp. 1–50.

[10] Voorhees, E.M.: Query expansion using lexical-semantic relations. In ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '94, Springer-Verlag New York, pp. 61–69 (1994),

[11] Abdou S. and Savoy J. Searching in Medline: Query expansion and manual indexing evaluation. Information Processing & Management, Volume 44, Issue 2, 2008, Pages 781-789.

[12] Sack, H.: Npbibsearch - an ontology augmented bibliographic search. In SWAP. CEUR Workshop Proceedings, vol. 166, 2005.

[13] Delfs, R., Doms, A., Kozlenkov, E., Schroeder, M..: Gopubmed: ontology-based literature search applied to geneontology and pubmed. In Proceedings of German Bioinformatics Conference. LNBI- Springer. pp. 169–178. (2004)

[14] Haase, P., Schnizler, B., Broekstra, J, et al..: Bibster a semantics-based bibliographic peer-to-peer system. In: Staab, S., Stuckenschmidt, H. (eds.) Semantic Web and Peer-to-Peer, Springer Berlin Heidelberg, 2006, pp. 349–363.

[15] Haslhofer B., Martins F. and Magalhães J. Using SKOS vocabularies for improving web search. In International Conference on World Wide Web companion (WWW '13 Companion). 2013, pp. 1253-1258.

[16] Savoy, J.: Bibliographic database access using free-text and controlled vocabulary: an evaluation. Information Processing & Management 41(4), 2005, pp. 873 – 890.

[17] Thesprasith, O. and Jaruskulchai, C.: Query expansion using medical subject headings terms in the biomedical documents. In Intelligent Information and Database Systems, LNCS, vol. 8397, Springer International Publishing, 2014, pp. 93–102.

[18] Beneventano, D., Bergamaschi, S., Guerra F. and Vincini, M.: The MOMIS approach to information integration. In: 3rd International Conference on Enterprise Information Systems - ICEIS - Setubal, Portugal, 2001, pp. 194–198

[19] Bromley J. King D and Morse, D. Finding agriculture among biodiversity: Metadata in practice. Communications in Computer and Information Science 478, 2014, pp. 185–192.

[20] Green JM, Harnsomburana J, Schaeffer ML, et al. Multi-source and ontology-based retrieval engine for maize mutant phenotypes. Database, 2011, pp. 1–15.

[21] Miller A. Wordnet: A lexical database for English. Communications of the ACM 38(11), 1995, pp. 39–41

[22] Salton G and Buckley C. Term-Weighting Approaches in Automatic Text Retrieval. Inf. Process. Manage. 24(5), 1988, pp. 513–523.

[23] Diaconis P and Graham RL. Spearman's footrule as a measure of disarray. Royal Statistical Society Series B 32(24), 1977. pp. 262–268.

[24] Marco AD and Navigli R. Clustering and diversifying web search results with graph-based word sense induction. Computational Linguistics 39(3), 2013, pp. 709–754

Appendix A

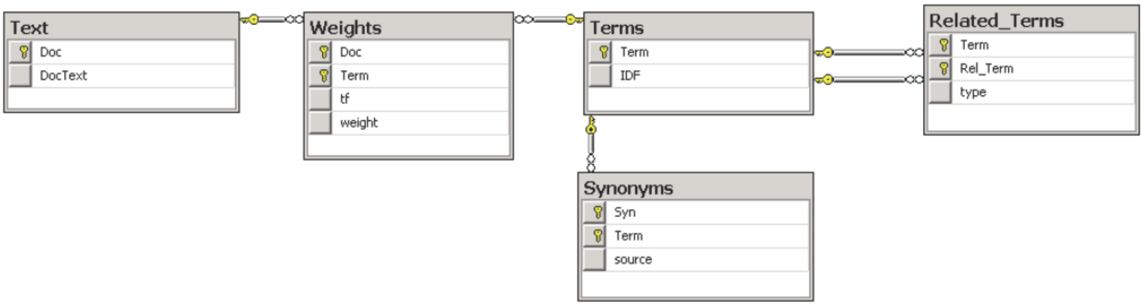


Figure 8. The semantic engine database schema.

The data structure of the Semantic Index provides a convenient way to store and access its data (Section 4.1). Data is maintained in a relational database (the schema is in Figure 8) is designed so to minimize the storage requirements and to provide good access efficiency. The tables of the schemas can be straightforwardly combined by simple SQL statements in order to extract the needed information (for instance to solve queries or to generate the conceptual views discussed in Section 4.1). Note that primary keys are <u>underlined</u>, while foreign keys are denoted with "FK". We also specify an indication on the column(s) on which to build the table indexes in order to maximize the querying performances.

Text (doc, docText)
Indexes on: (doc)

This table contains the documents that have been analyzed, and includes a “doc" key (i.e. the codename of the document, such as “D00001”) and the document text itself (“docText").

Terms (term, idf)
Indexes on: (term)

The “Terms” table contains the terms as extracted from the documents, which include a “term” key (i.e. the term itself such as “frost resistance”) and the inverted document frequency (“idf”) of the term in the document collection.

Weights (doc, term, tf, weight)
Indexes on: (doc, term), (term)
FK: doc references Text(doc)
FK: term references Terms(term)

This table contains the weights of the terms. In particular, given a document “doc” from the Text table and a term “term” from the Terms table, “tf” represents the term frequency of such term in such document and “weight” represents the product of the term frequency and inverse document frequency. An index on the single “term” column also helps in quickly identifying all the documents in which a given term appears without keeping this information in a dedicated column.

Synonyms (syn, term, source)
Indexes on: (syn)
FK: term references Terms(term)

It contains the synonyms of the terms. Instead of explicitly have an entry in “Terms” for all the different synonyms of a term, the idea is to choose one representative term to be inserted in the “Terms” table, and for each alternative “syn”, to store in “Synonyms” the term(s) of which “syn” is a synonym. “source” encodes the knowledge source providing the synonym (i.e. AGROVOC or WordNet).

Related_Terms (term, rel_term, type)
Indexes on: (term, rel_term), (term)
FK: term references Terms(term)
FK: rel_term references Terms(term)

This table encodes the semantic relatedness relation, including narrower and correlated terms (“type”=‘N’ and ‘C’, respectively): for each term “term", the table indicates the term(s) “rel_term” to which “term” is semantically related.