# A User-aware and Semantic Approach for Enterprise Search

**Giacomo Cabri**

FIM - University of Modena and Reggio Emilia
Via Campi 213/b, 41125, Modena, Italy
Tel. +39 059 205 8320, Fax +39 059 205 5216
giacomo.cabri@unimore.it

**Riccardo Martoglia***

FIM - University of Modena and Reggio Emilia
Via Campi 213/b, 41125, Modena, Italy
Tel. +39 059 205 8322, Fax +39 059 205 5216
riccardo.martoglia@unimore.it

### Abstract

In addition to general purposes search engines, specialized search engines have appeared and have gained their part of the market. An *enterprise* search engine enables the search inside the enterprise information, mainly web pages but also other kinds of documents; the search is performed by people inside the enterprise or by customers. This paper proposes an enterprise search engine called AMBIT[i]-SE that relies on two enhancements: first, it is user-aware in the sense that it takes into consideration the profile of the users that perform the query; second, it exploits semantic techniques to consider not only exact matches but also synonyms and related terms. It performs two main activities: (i) information processing to analyse the documents and build the user profile and (ii) search and retrieval to search for information that matches user's query and profile. An experimental evaluation of the proposed approach is performed on different real websites, showing its benefits over other well-established approaches.

***Keywords***— User-awareness, Enterprise Search Engine, Information Retrieval, Text Analysis, Semantic Knowledge and Similarity.

## INTRODUCTION

Enterprises produce and rely on a large amount of information. A small part of information is available by public web sites, while the most part is exploited by *employees* of the enterprise itself by means of intranet, and by *customers* of the enterprise who have access to some information for business purposes.

In this scenario, the capability of searching for needed information plays a fundamental role. On the one hand, enabling internal employees to find the needed information in a short time is not only useful to speed up their work, but also to avoid or decrease the frustration of long and unsuccessful searches. On the other hand, precise and relevant answers to customers that exploit the company web sites for both searching and interacting can grant a high degree of customer satisfaction.

In this scenario, the authors point out two aspects that can improve the use of search engines in an enterprise context by providing more relevant search results: *user-awareness* and *semantics*. Existing studies and surveys in general information management contexts have highlighted the benefits that can be brought to search results by the former (Xiang et al., 2010) and latter (Mangold, 2007). User-awareness means to exploit the knowledge of the user in terms of profile and context to effectively tailor the search on the base of the available information. Semantics can be useful to overcome the limitations of a syntactic approach, which is often exploited but does not consider similar pieces of information expressed in different ways. As far as the authors know, there are no enterprise search engines that exploit both aspects in a single approach.

Starting from the above considerations, the general semantic foundations introduced in (Martoglia, 2015) are exploited to proceed towards the goal of achieving a user-aware semantic enterprise search engine. The search engine is called AMBIT-SE (AMBIT Search Engine). It is not a generic search engine, but a search engine dedicated to searches in an enterprise scenario. The AMBIT-SE approach improves search results by:

- taking advantage of textual information, including user information. Indeed, text is the primary component of the documents that should be presented / suggested to users, and also one of the main information characterizing user profiles. Consider, for instance, the contents of user browsing history, the description of users' interests, and so on;
- exploiting semantic techniques: instead of a pure syntactic matching between the query keywords and the words in the available documents, it relies on their meaning and takes into account synonyms and related terms.

The approach presented in this paper brings the following novel contributions with regard to the initial idea of an enterprise search engine sketched in (Cabri, 2016) and to the state of the art:

- differently from the state of the art on available enterprise search engines, it is able to combine semantics and user-awareness without requiring any manual work (e.g. for annotating documents, describing user profiles, etc.). Novel semantic and user-aware techniques allow the engine to go beyond standard syntactic search in a completely automatic way;
- the *semantic text analysis* techniques are adapted and refined from previous authors' studies on the effectiveness of semantic text management in specific subject areas such as software engineering (Bergamaschi et al., 2015; Martoglia, 2011), agricultural (Beneventano et al., 2016) and user-centric cultural enhancement data (Martoglia, 2015). In the presented approach, the techniques are generalized to work in a non-specialized enterprise search setting with general purpose ontologies and new weighting schemes, allowing them to be directly compared to the new class similarity contribution;
- the strength of the semantic text analysis contribution is added to the new contribution given by *semantic categorization*. Categorization classifies documents on the basis of a well-known taxonomy (defined by IPTC[ii]) in order to provide improvements in the retrieval effectiveness. To this end, a novel class similarity metric is introduced, with a weighting scheme exploiting the novel concept of *inverse (document) class frequency*;
- a novel ranking selection/fusion technique is employed, producing a final document ranking which: (a) reflects both the query and user profile in a flexible and

customizable proportion; (b) fuses both class and text similarity contributions in the case they are judged as significant; (c) otherwise, it is able to automatically exclude one of the two contributions from the final result.

The combined *text analysis*, *user-aware* and *semantic retrieval* techniques ultimately provide enhanced searching effectiveness over standard search techniques, as also shown in the experimental tests. Moreover, the approach is devised for IT Small and Medium-sized Enterprises (SMEs), providing them with easy-to-apply methods that allow them to query for the information they need in the way they are used to.

This paper is organized as follows. First, a related work analysis is presented (Section "Related Work"). Then, Section "Search Engine Architecture" describes the architecture of the proposed search engine. Section "Information Processing" explains how the presented approach processes the documents that can be "searchable" by the users. Section "Search and Retrieval" illustrates how the system defines a ranking of retrieved documents to satisfy the user's query. Finally, the results of the experiments carried out on the system are reported, before the conclusions (Section "Conclusions"), in Section "Experimental Evaluation".

# RELATED WORK

This section presents a review of the existing approaches related to user-aware semantic enterprise search engines. The authors have considered both academic and commercial approaches, and have classified them on the base of two aspects: the *user-awareness* and the *semantics*; the resulting taxonomy is reported in Figure 1.

Most of the approaches do not consider together these aspects and/or cannot be classified as enterprise search engines. For these reasons, related work will be presented in three separate subsections: semantic approaches, user-aware approaches and enterprise search engines.

Figure 1: Taxonomy of user-aware semantic approaches

## Semantic Approaches

The Semantic Web is likely to be the field where most of the approaches for semantic document retrieval has been proposed, as reported in (Mangold, 2007). It is a survey which covers approaches that exploit domain knowledge to process search requests. The authors discuss a large variety of domain knowledge utilization that include automatic query expansion and ontology-driven document retrieval.

The possible ineffectiveness of information retrieval systems is mainly due to the inaccuracy with which a query formed by a few keywords models the actual user information need. One well known method to solve this problem is automatic query expansion, whereby the user's original query is augmented by new features with a similar meaning (Carpineto & Romano, 2012). Differently from the approach presented in this paper, complex query expansion techniques, such as the ones discussed, usually require different parameters to be specified (as also stated in (Abdou & Savoy, 2008)). For instance, the method proposed in (Voorhees, 1994) involves a set of parameters specifying for each run and for each relation type included in the ontology 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 (Abdou & Savoy, 2008) and therefore a long process of manual tuning is needed.

More and more document retrieval systems make use of ontologies to help users better specify their information needs and produce semantic representations of documents. (Haslhofer et al., 2013) proposes a Simple Knowledge Organization System (SKOS) based term expansion and scoring technique that leverages labels and semantic relationships of SKOS concept definitions. The Hybrid spreading activation approach (Rocha et al., 2004) requires tight coupling between the document base and the ontology, which is a graph where concepts and properties are nodes and edges, respectively. The set of nodes that match the given query terms is used as the start nodes of a spreading activation algorithm: documents with highest activation are ranked highest in the result set. Differently from the presented approach, the mentioned systems have no notion of user context.

Given the need for manual intervention, typical semantic retrieval techniques obtain a good degree of effectiveness only on manually annotated collections and/or with explicit user intervention. (Savoy, 2005) compare the retrieval effectiveness of different search models in a bibliographic database context. These models are founded on automatic syntactic text-word indexing or on manually assigned controlled descriptors. (Thesprasith & Jaruskulchai, 2014) proposes a query expansion technique working on MEDLINE documents that have been manually assigned to controlled MeSH (Medical Subject Headings) vocabularies. The indexing and retrieval approach proposed by AMBIT-SE, instead, exploits the semantics of the text while remaining completely automatic.

(De Vocht et al., 2017) presents a semantic search engine focusing in particular on integration of different sources of data in the science research field. To increase the precision of the results, the authors annotated and interlinked structured research data with ontologies from various repositories exploiting a semantic model. That approach does not consider user-awareness and requires annotation.

(Figueroa & Neumann, 2016) focuses on the search in the context of Community question answering (cQA) platforms. It induces the semantic classes of question-like search queries by means of the contextual information. Context is set up or represented by inferred views of their respective search sessions, namely views modelling previous queries entered by the same user. The idea of introducing semantic classes and of combining them with contextual information is very interesting, but in that work is limited to a specific field (cQA).

SINA (Shekarpour et al., 2015) is a scalable keyword search system that can answer user queries by transforming user-supplied keywords or natural-languages queries into conjunctive SPARQL queries over a set of interlinked data sources. In this case, differently from the scenario considered in this paper, data are expressed in graph format. The system exploits semantics to improve search over different data sources, but does not take into consideration user information to refine queries.

## User-Aware Approaches

The advantages of taking into consideration the context has been point out several times in the literature (Bolchini et al., 2011; Cabri et al., 2003; Falcarin et al., 2013; Xiang et al., 2010; Vu et al., 2017). In particular, a few works concern context modelling, representation, and effective handling. For instance, (Bolchini et al., 2011) proposes to design a context management system which is not application-dependent, (Falcarin et al., 2013) proposes an architectural framework for context data management, while (Villegas & Müller 2010) reports the result of a study on various context modelling and management approaches. (Xiang et al., 2010) addresses the problem of integrating context information into a ranking model. (Liu et al., 2004) proposes a method to derive a user profile based on the search history and on pre-determined category hierarchies. (Vu et al., 2017) proposes a personalised query suggestion framework for Intranet search, relying on two temporal user profiles.

Most of these approaches primarily focus on specific aspects such as external conditions or location, they do not consider the semantics of the context and/or they rely on manual work in order to classify and categorize users and documents. General purpose search engines, such as Google, typically provide only very simple localization of search results on the base of the IP-address.

## Enterprise Search Engines

Many enterprise search engines have been proposed in the literature or are sold by companies. Most of the proposals do not provide an ontology-based *semantic* analysis, relying instead on *syntactic* and *hand-coded* rules. Some examples are Alfresco[iii], Autonomy[iv], Solr[v]. Google[vi] is also an example of a syntactic search engine that can be exploited in a enterprise search context.

There are, of course, exceptions to this rule. The SHOE project (Heflin & Hendler, 2000), which requires a domain-ontology where document types correspond to ontology concepts. Attivio[vii], a product which manages information in the RDF format, providing search results and alerts. Expert System's Cogito[viii], which provides automated disambiguation, classification, entity extraction, and metadata. Nevertheless, these systems provide no notion of user context. Instead, Coveo[ix] is a tool specifically oriented to exploit contextual knowledge for dealing with information related to customers and agents, but it does not exploit semantic information.

Considering the semantic and context information, there are few systems that exploit it, even if in a sometimes limited way. The Ontogator system (Hyvonen et al., 2003), part of an image management and retrieval system, provides an interactive recommendation system that allows the user to browse images based on ontological properties. To exploit user contexts, it introduces views to the ontology that rely on different concept hierarchies, called "facets". Each view represents a specific information-need. PrEmISES, proposed in (Ramona-Cristina et al., 2016), is a framework that aims at addressing information management needs of SMEs relying on ontologies to add semantically enabled information integration. The framework definition is still in a very preliminary phase; therefore, semantic and context-sensitive features are currently only sketched. Another example is IBM's Content Analytics with Enterprise Search[x], which exploits a framework called Unstructured Information Management Architecture (UIMA), in order to build analytic applications and to find meanings, relationships and relevant facts hidden in unstructured text. Context information is provided by means of manual annotations. These approaches require manual intervention on the documents and/or adopt a still limited notion of context, i.e. they do not exploit all of the data potentially available related to the user, such as the contents of any web page visited, attachment downloaded, and similar documents.

## Discussion

The authors have reported several researches that are connected to the approach presented in this paper. They point out that there is no existing approach that addresses all the following aspects at the same time:

- *Semantics*: exploitation of semantic techniques to improve the results of the query;
- *User-awareness*: exploitation of user information to customize the results of the queries;
- *Generality*: capability of being applied to general sources of information, not to a specific field;

- *Automation*: no need for manual annotation of the information.

Starting from the above analysis, the AMBIT-SE approach aims at addressing all the above-mentioned aspects. In the next section, the architecture of AMBIT-SE is presented.

# SEARCH ENGINE ARCHITECTURE

Figure 2: An overview of the AMBIT-SE architecture

Figure 2 shows the architecture of the proposed search engine, the main activities and the related modules. In (Cabri et al., 2016) the authors tackled in detail many text pre-processing issues such crawling techniques, paragraph identification and text tagging. In this paper, the focus is on defining the *semantic text analysis* and the novel *semantic categorization* techniques (described in Section "Information Processing") The general architecture is also significantly extended in order to manage the output of both kinds of techniques and to exploit it in the actual search (Section "Search and Retrieval"). The architecture highlights two main activities:

1. **Information processing** (dashed line in figure). This activity is applied to both the documents to be retrieved (e.g. web pages for a given site) and the documents useful to determine the user's profile (such as e-mails, web pages viewed, profile information, past search queries, etc.). The available information is extracted (crawled) and indexed by means of ad-hoc techniques, also exploiting external knowledge sources. The data structures containing all the information processing results will be referred to as "semantic glossaries", one for the website(s) and one for the user profile. Both glossaries are then compared ("Semantic glossaries computation and comparison" in the figure) with ad-hoc document similarity algorithms, detailed in Section "Search and Retrieval". The generated "Profile rankings" symbolize how relevant the retrievable documents are in relation to the user's profile.
2. **Search and retrieval** (solid line in figure). This activity provides useful answers to the user by retrieving the most relevant documents with regard to the user query, also taking into account its pre-computed profile information. This is achieved ("Semantic query processing") by determining the "Query rankings" of the query with regard to the available documents (semantic similarity discussed in Section "Search and Retrieval"). Query rankings are finally fused with profile rankings in order to provide the final ranking.

These two main activities are detailed in the next two sections.

# INFORMATION PROCESSING

Information processing is common to both the documents of the website(s) to be indexed and the user profile documents. Each of them will contribute to a semantic glossary structure containing the results of the analysis.

The first step of the information processing is *crawling*: web/file crawling is performed in order to retrieve the raw data (e.g., Title, Content, URL, File Name, Meta Description, Meta Keywords) of the documents that will be further analysed by the subsequent steps. All of the crawling operations are handled by the open-source enterprise-class search engine software, OpenSearchServer[xi].

## Semantic Text Analysis

As single existing tools do not allow sufficient configuration and extension options, the authors designed a customized *Semantic text analysis* module, which exploits several open source libraries to perform specific actions. The analyser allows us to extract the contents (and meanings) of the processed information, by means of:

- *Text extraction*, an operation that can vary greatly depending on the format of each document; this is taken care of by components of the open-source software GATE[xii];
- *Tokenization*, the terms are identified and punctuation is removed;
- *Stemming and Part Of Speech (POS) Tagging*, the tokens are "normalized" and "stemmed", i.e., they are reduced to their base form (managing plurals and inflections) and "tagged" with POS tags (i.e., nouns, verbs, ...); this is taken care of by the TreeTagger[xiii] library;
- *Composite term identification*, possible composite terms (such as "production area" or "wine tasting") are identified by means of a simple state machine and of POS tags information;
- *Word Sense Disambiguation (WSD)*, the meaning of the keywords resulting from the previous steps is made explicit with regard to a reference thesaurus (typically, WordNet[xiv], (Miller, 1995)); in particular, the relevant synsets are extracted and associated;
- *Weight computation*, keywords are finally enriched with weight information (see below).

The extracted information enables the retrieval of the most relevant information for the user in the search and retrieval phase. In particular, the proposed approach exploites the semantic information from the thesaurus and the similarity functions described in Section "Search and Retrieval". Thanks to them, AMBIT-SE will be able to retrieve documents on the basis of synonyms and related keywords; for instance, documents about "dogs" are considered as relevant to a query about a "terrier". Moreover, by means of WSD, documents about a "pet" in the sense of "domesticated animal" will not be mixed up with those where "pet" means "a special loved one".

For the weight information, text analysis exploits a variant of the standard "tf-idf" weighting scheme. This is introduced to convey the information of the classic scheme while also keeping the weights normalized in the range of [0,1]: this enables an effective ranking comparison and fusion (more on this in Section "Search and Retrieval"). Given a document $D^x$, each keyword $k_i^x \in D^x$ is assigned a *keyword weight* $kw_i^x$ defined as:

$$kw_i^x = \overline{kf}_i^x \cdot \overline{idf}_i \qquad (1)$$

where:

$$\overline{kf}_i^x = \frac{f_i^x}{max_l f_l^x} \qquad (2)$$

$$\overline{idf}_i = log\left(\frac{N}{n_i}\right) / \; max\,_l(log\left(\frac{N}{n_l}\right)) \qquad (3)$$

$\overline{kf}$ and $\overline{idf}$ are *normalized* keyword frequency and *inverse* document frequency: $f_i^x$ is the raw frequency of keyword $k_i$ in document $D^x$, *N* is the total number of indexed documents and $n_i$ the number of documents where keyword $k_i$ appears. Note that, by definition, $0 \leq \overline{kf}_i^x \leq 1$ and $0 \leq \overline{idf}_i \leq 1$.

## Semantic Categorization

Besides semantic text analysis, a *semantic categorization* process is also applied to the documents in order to tag each document with appropriate subject classes. The Media Topic NewsCodes taxonomies and vocabularies provided by IPTC are adopted. These are well-known taxonomies offering a very good level of detail and coverage of a wide range of topics. In this case, the starting point is the output of ad-hoc categorization tools from Expert System[xv], where each class tag is given a score $s(c_i)$, $0 \leq s(c_i) \leq 1$; the higher the weight the more relevant the class is for the document. For instance, a document about the typical "Terrier food products" will presumably have "Pet product and service" among its highest scoring associated tags. However, the authors deem that if most of the documents of the collection are tagged with the same class, this class will not be particularly distinctive and useful in the retrieval phase. Therefore, this fact is captured by going beyond the plain score and by defining a new weighting scheme for the classes inspired by the text analysis scheme (Eq. 1): given a document $D^x$, each class tag $c_i^x \in D^x$ is assigned a *class weight* $cw_i^x$ defined as:

$$cw_i^x = s(c_i^x) \cdot \overline{icf}_i \qquad (4)$$

where:

$$\overline{icf}_i = log\left(\frac{N}{t_i}\right) / \, max_l(\log\left(\frac{N}{t_l}\right)) \qquad (5)$$

$\overline{icf}$ stands for *inverse (document) class frequency*; *N* is the total number of documents and $t_i$ is the number of documents tagged with class $c_i$.

## Website(s) and User Semantic Glossaries

By applying batch information processing to the document collection, the *website(s) semantic glossary* is automatically generated. Conceptually, it consists of a *global view* (all keywords/classes together with their occurrences and additional extracted data), and a *per-document view* (keywords/classes occurrences in each document with their statistics). A simplified sample of the latter view is shown in Table 1-left and Table 1-right). In particular, "Document" is the list of the documents IDs in which each keyword/class occurs, "Synsets(s)" are the synsets selected as the output of WSD, while "KF", "KWeight" and "CWeight" are the weights illustrated in Eqs. 1 and 2.

| **Document** | **Keyword** | **Synset(s)** | **KF** | **KWeight** |
|---|---|---|---|---|
| P02001 | Terrier | 00919240-n | 0.445 | 0.277 |
| P02005 | Veal | 00414222-n | 0.210 | 0.131 |
| … | … | … | … | … |

| **Document** | **Class** | **CWeight** |
|---|---|---|
| P02001 | Pet product and service | 0.645 |
| P02005 | Food industry | 0.442 |
| … | … | … |

Table 1: Sample portions of the extracted Document Semantic Glossary: per-doc view for keywords (left) and classes (right).

Each time a user logs in AMBIT-SE, a batch analysis is also scheduled on the data of her profile *U* in order to generate/update its *user semantic glossary* (which shares the same structure of the website semantic glossary discussed above). In particular, the profile contains *action history data* including a list of past accessed documents and past searches performed on the website. The idea is that such data can be analysed in the same way as the website documents, therefore exploiting all the power of the document text analysis and categorization in order to associate meaningful keywords and classes to users; these will provide more for relevant results to their queries in the search phase. Due to their complexity and so the need for computational power, all the analyses are performed when no users are logged in; they will be available in order to process future requests from the same user in a more accurate way.

# SEARCH AND RETRIEVAL

When a user *U* submits a query *Q*, AMBIT-SE has to answer it as effectively as it can. The technique illustrated in this section aims to produce the most effective ranking $\hat{\tau}$ of the indexed documents $D \in \mathcal{D}$with regard to both *U* and *Q*. This is done by taking into account all the information available in the semantic glossaries produced by the analysis process (text analysis and classification) and by means of the following ad-hoc similarity metrics between two generic documents (i.e. keyword sets) $D^x$ and $D^y$:

- the similarity $TextSim(D^x, D^y)$, considering *keyword* information;
- the similarity $ClassSim(D^x, D^y)$, considering *class* information.

In particular, *TextSim* and *ClassSim* will be applied to both:

- the user profile *U* (i.e., $D^x=U$) with regard to each indexed document $D^y=D\in\mathcal{D}$(as performed on past navigated data): this produces the profile rankings $\tau^U_{text}$ and $\tau^U_{class}$;
- the query *Q* (i.e., $D^x=Q$) with regard to each document $D^y=D\in\mathcal{D}$, this produces the query rankings $\tau^Q_{text}$ and $\tau^Q_{class}$.

In the final part of this section, the authors will show how this ranking information is exploited in *ranking selection/fusion* in order to produce the final ranking $\hat{\tau}$. Next, the *TextSim* and *ClassSim* similarity metrics will be defined.

## Text and Class Similarity Metrics

Building on previous research on text retrieval for specific subject areas as software engineering (Bergamaschi et al., 2015; Martoglia, 2011), agricultural (Beneventano et al., 2016) and user-centric cultural enhancement data (Martoglia, 2015), the text similarity $TextSim(D^x, D^y)$ formula between documents $D^x$ and $D^y$ is defined as:

$$TextSim(D^x, D^y) = \frac{\sum_{k_i^x \in D^x} max_{k_j^y \in D^y} (KSim\ (k_i^x, k_j^y)) \cdot kw_i^x \cdot kw_j^y}{|D^x|}\ , \qquad (6)$$

where the weighting scheme illustrated in Eq. 1 is exploited and *KSim* is a term similarity formula (see later) taking into account the semantic information extracted from the semantic glossary. Simply put, the similarity $TextSim(D^x, D^y)$ between two documents $D^x$ and $D^y$ is determined by summing the maximum keyword similarity scores $KSim(k_i, k_j)$ between each pair of keywords $k_i$ and $k_j$ belonging to different documents, multiplied by the weights of both. As to $KSim(k_i, k_j)$, in AMBIT-SE semantic framework, $k_i$ and $k_j$ can be:

1. *equal or synonyms*: $KSim( , )$=1;

2. *related*, i.e. the thesaurus hypernymy path similarity between the keywords'synsets exceeds a given threshold *Th* (Bergamaschi et al., 2015): $KSim(\,,)$=*r*, where *r* is an arbitrary value between 0 and 1, default 0.7;

3. *unrelated* otherwise: $KSim(\,,)$=0.

Please note that, since both $0 \leq KSim(\,,) \leq 1$ and $0 \leq kw \leq 1$, then $0 \leq TextSim(\,,) \leq 1$. This will be useful in the ranking selection/fusion phase (Section "Ranking Selection / Fusion").

In addition to document keywords, the classes associated by the semantic classifier can also significantly help in retrieving useful documents. This is obviously true if both documents are strongly characterized by a common IPTC class (e.g. "process industry"); however, also documents about cola factories tagged with a similar class "food industry" would be of interest. This is achieved through $ClassSim(D^x, D^y)$, which quantifies the similarity of $D^x$ and $D^y$ on the basis of their associated IPTC classes $c_i^x \in D^x$ and $c_j^y \in D^y$:

$$ClassSim(D^x, D^y) = \frac{\sum_{c_i^x \in D^x} max_{c_j^y \in D^y} (CSim\,(c_i^x, c_j^y)) \cdot cw_i^x \cdot cw_j^y}{|\{c^x \in D^x\}|} \qquad (7)$$

As for Eq. 6, $CSim(c_i, c_j)$ between classes $c_i$ and $c_j$ ranges between 1 (equal classes), *r* (similar, i.e. "near" on the IPTC taxonomy, classes) and 0 (otherwise). Again, since $0 \leq CSim(\,,) \leq 1$, then $0 \leq ClassSim(\,,) \leq 1$.

# Ranking Selection/Fusion

As previously seen, given a profile *U*, Eqs. 3 and 4 induce rankings $\tau_{text}^U$ and $\tau_{class}^U$ on each document $D \in \mathcal{D}$with regard to *U*, respectively. Similarly, when the two equations are computed with regard to a query *Q*, rankings $\tau_{text}^Q$ and $\tau_{class}^Q$ are induced on the same documents. The aim of the *ranking selection/fusion* process presented in this paper is to exploit the information of those rankings in order to obtain the final ranking $\hat{\tau}$. The aims are the following:

1. the final ranking $\hat{\tau}$ should reflect both *Q* and *U*, and it should be possible to define the relevant importance of the query *Q* with regard to the profile *U* (for instance, to privilege the information contained in *Q*);

2. the final ranking $\hat{\tau}$ should be flexibly defined so to reflect both text and class information (Eqs. 3 and 4, respectively), possibly in a customizable proportion. Moreover, the system should be able to automatically decide if one of the two kinds of rankings (text, class) is not significant, for instance due to very low similarities, which would only negatively affect the final ranking, therefore excluding it from the fusion;

3. the *score* of each document *D*, and not only its position in the rankings, should be taken into account in order to directly reflect its relevance in the final ranking.

Let us see how AMBIT-SE achieves these points. By means of a linear combination score fusion method, the scores inducing the fused rankings $\tau_{text}$ and $\tau_{class}$, which take into account both *U* and *Q*, are defined as:

$$s^{\tau_{text}}(D) = \alpha_Q \cdot s^{\tau_{text}^Q}(D) + (1 - \alpha_Q) \cdot s^{\tau_{text}^U}(D), \qquad (8)$$

$$s^{\tau_{class}}(D) = \alpha_Q \cdot s^{\tau_{class}^Q}(D) + \left(1 - \alpha_Q\right) \cdot s^{\tau_{class}^U}(D),$$

where $0 \leq \alpha_Q \leq 1$ is a preference weight determining the relevant importance of $Q$ with regard to $U$. The default value is $\alpha_Q$= 0.7, meaning that the information in $Q$ is typically more significant than that in $U$; this can be varied, for instance, by the system administrator depending on the actual usage scenarios.

The above defined $\tau_{text}$ and $\tau_{class}$ rankings are eventually fused in a final ranking $\hat{\tau}$, which takes into account both text and class contributions as of Eq. 8:

$$s^{\hat{\tau}}(D) = \beta_{text} \cdot s^{\tau_{text}}(D) + (1 - \beta_{text}) \cdot s^{\tau_{class}}(D) \qquad (9)$$

where $0 \leq \beta_{text} \leq 1$ is a preference weight determining the relative importance of $\tau_{text}$ with regard to $\tau_{class}$. The default value is 0.5. Note that ranking selection is implemented in the following way: if $\sum_D s^{\tau_{text}}(D) \gg \sum_D s^{\tau_{class}}(D)$ then $\beta_{text}$ is automatically set to 1 in order to exclude the contribution of $\tau_{class}$ (and vice-versa), therefore avoiding possibly detrimental noise. Only documents that are part of all rankings will appear in the final ranking.

# EXPERIMENTAL EVALUATION

| Website | Description | Examples of information needs |
|---|---|---|
| http://www.cobat.it/ | A relatively small website that provides information and services for disposing and recycling four problematic waste categories: batteries and accumulators, tires, electric and electronic devices, and photovoltaic panels. | Retrieve documents pertaining to the disposal of batteries and accumulators. |
| http://evergreensmallbusiness.com/ | A Blog that publishes different kinds of information and advice for small businesses, all classified in categories such as business taxes, management, personal finance, etc. | Retrieve articles pertaining to bookkeeping. |
| http://www.bagnolottanta.it/ | A journalistic website with several thematic columns such as history, culture, theatre, etc. | Retrieve articles in specific columns such as books. |
| http://truegoods.com/ | An Indie online shop that specializes on healthy and natural products. | Retrieve information on products belonging to the pet-care category. |
| http://www.gruppozatti.it/ | An authorized car dealer which sells several brands of both new and used cars. | Retrieve different information about cars belonging to the used category. |

Table 2: Experimental setting: details of the five business-relevant websites selected for evaluation purposes.

This section presents the results of several tests performed on different kinds of websites. This paper is focused on effectiveness evaluation; for readers interested in time performances, the current prototype has a response time of 40 ms on average on a standard single-node configuration.

Five websites were selected for evaluation purposes; for each one of them, appropriate information needs were established by examining common searches performed in the past.

The choice of the website was driven by the aims of the proposed approach. First, the sites are all representative of *real* and typical *business-relevant* sites. Moreover, the sites were selected so as to cover a number of *different topics* (e.g. health, cars, recycling, …) and thus different terminologies and text content. Finally, they have different features in terms of structure. This is in line with the main goal of this section, i.e. to evaluate the *effectiveness* and the *flexibility* of an enterprise search engine.

As to possible experimental comparisons, as anticipated in the related, there are currently no approaches that can be directly compared with the proposed one. In particular, among the approaches whose implementation was publicly available, those offering semantic features require unavailable manual annotations and/or were strictly designed to work on specific ontologies (Haslhofer et al., 2013; Thesprasith & Jaruskulchai, 2014; De Vocht et al., 2017). This is also true for semantic enterprise search engines (Cogito, Attivio, Content Analytics). The considered user-aware search engines do not exploit the semantics (Bolchini et al., 2011; Cabri et al., 2003; Falcarin et al., 2013; Xiang et al., 2010; Vu et al., 2017). Also, the publicly available approaches working on context information are restricted to aspects such as time and location (Falcarin et al., 2013; Xiang et al., 2010; Villegas & Müller 2010) or very specific scenarios (e.g. agent management for Coveo, image search for Ontogator, cQA for (Figueroa & Neumann, 2016)) which makes them not applicable to the considered case. Summarizing, the reason that prevents direct comparisons are: first, only a few number of approaches exploit both semantics and user awareness; second, most of the approaches focus on a specific application field and cannot be applied to general websites; third, several approaches are not freely available to be tested. Anyway, for reference, the final part of this section is devoted to a direct comparison with Google search engine; even if it does not exhibit all the features of AMBIT-SE, the authors considered it an important benchmark in the field.

Table 2 reports, for each website, the address, a brief description of its contents and an example of a considered information need. In the context of the information needs of each website, 50 plausible queries were submitted to the system. Moreover, two options for user profiles are considered: a simpler setting, where the profile contains only documents relevant to the information need ("homogeneous profile"), and a more complex one, where the profile contains also an equal number of irrelevant documents ("heterogeneous profile").

In each situation, the output of AMBIT-SE is compared with a "gold standard", i.e. relevant answers manually selected from the websites, and precision and recall are assessed; in particular, the tests compute interpolated precision at 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0 (0, 10, 20, ..., 100 percent) (Baeza-Yates & Ribeiro-Neto, 1999), thus taking into account not only the relevance of the results but also their position in the returned ranking. Moreover, a "stable" situation is assumed, where users and documents have been already automatically processed and their relevant keywords and classes stored in the Semantic Glossaries. All the parameters are set at default.

Figures 3-5 depict the results (averaged on all the queries) obtained by AMBIT-SE, compared with the baseline of a syntactic retrieval method ignoring synonyms, related terms and class information. In particular, this baseline is representative of the document retrieval techniques commonly exploited by most commercial systems and standard enterprise search engines (e.g. Alfresco, Autonomy, Solr). Further comparisons with available systems (e.g. Google) will be discussed in the final part of the experimental analysis. In order to better visualize the difference between AMBIT-SE's results and the baseline, the figures detail the the precision achieved by the different features of AMBIT-SE. In particular, the white bars (at the bottom) represent the precision of the syntactic method (baseline); the contributions on top are the improvements achieved by means of the semantic (black) and user-aware features (gray).

Figure 3: Test results for *http://www.cobat.it/* (left) and *http://evergreensmallbusiness.com/* (right)

As readers can see, the improvements in precision offered by the semantic and user-aware features of AMBIT-SE are significant in each scenario, ranging from 20% to even 80%. Let us start by analysing the *http://www.cobat.it/* results (Figure 3 (left)). Here, the semantic features offer an advantage at all recall levels; the results also benefited from the user-aware features, because of the large number of keywords contained within the documents associated to the user profiles. For instance, the use of semantics enables the matching of different but very related terms like “battery” and “accumulator”. Those contributions go unnoticed in standard syntactic search.

Considering the second website (Figure 3 (right)), the use of semantics provides even greater benefits: for instance, the use of synonyms and related terms are able to exploit a number of terms correlations such as between “money” and “bookkeeping”. This was especially evident in the longer and less direct queries submitted, which are harder to satisfy by a search engine without additional information in the form of synonyms and related terms. Summing up the contribution of the profile analysis, AMBIT-SE reaches in some cases an improvement of nearly 80% in precision.

Figure 4: Test results for *http://www.bagnolottanta.it/* (left) and *http://truegoods.com/* (right)

In the test of Figure 4 (left), the authors found that in some cases the queries did not directly benefit from synonyms and related terms management, maybe due to very specific terminology that is used in the pages. Anyway, the use of semantics and the profile analysis provide relevant advantages even in this scenario. Moving to Figure 4 (right), the use of word stems, synonyms, related terms and profile can turn even difficult queries into manageable ones. For instance, in some cases, the syntactic baseline could not retrieve any records since the queries don’t include terms that match exactly the ones found in the relevant documents: among the exploited terms correlations, the very frequent one between “pet” and “animal”.

Figure 5: Test results for *http://www.gruppozatti.it/*

The final website considered (Figure 5) is characterized by pages containing very little text, thus providing a different task with regard to the others. Also in this case, a lot of complex queries did not yield satisfactory results for the syntactic baseline, thus providing very consistent advantages from the semantic features (precision improvement of 30% or greater). Due to the peculiar nature of the website, some queries were apparently too difficult even with the additional input provided by the semantics and profiles. Anyway, on average, the effect of the advanced AMBIT-SE features is evident even in this situation.

Figure 6: Class improvement analysis (left) and comparison with Google (right)

A further test is presented which is aimed at evaluating the impact of the classification features of the presented engine, including the novel ranking selection and fusion capabilities. Figure 6 (left) shows the precision figures (averaged over all the queries and all recall levels) that are achieved by exploiting only the text ranking and similarities (left bars) with regard to the ones achieved by the complete setup presented in this paper. The benefits are usually in the order of more than 10% of improvement. Looking at specific cases, the fused ranking was able to take the best from the class and text rankings, together capturing the user interests more completely.

In order to complete the evaluation, the authors performed a quantitative comparison with a state-of-the-art search engine and a discussion about results is reported in the following. The chosen search engine was Google, as mentioned before.

Figure 6 (right) shows the results of the system compared with those obtainable through the Google search engine. Indeed, Google is one of the most widely used search engines also in the enterprise search area. Similarly to AMBIT-SE, Google is completely automatic and requires no manual work (e.g. annotation) on the documents to be processed. In this case, the considered document set has been restricted to the specific considered website pages. Differently from AMBIT-SE, Google does not perform semantic analysis of the texts, however it employs text processing techniques such as stemming. Multiple users were simulated in Google through different search sessions. The results show that the average precision achieved by Google (in the range of 10-30%) is quite lower than the one achieved by AMBIT-SE (always above 55%). In particular, Google offers a level of performance that is only marginally better than the syntactic baseline discussed in the previous tests.

# CONCLUSIONS

Search engines represent a means essential for a lot of activities. This is especially true in an enterprise context, where the success of the enterprise is often strictly dependent on the ability of its employees and customers to find the needed information.

In this context the authors have proposed AMBIT-SE, an enterprise search engine approach that relies on two aspects, *user-awareness* and *semantics*, jointly exploited. First, it builds a profile of the user, which is exploited to search for the information that best matches with her needs. Second, semantic techniques enable the retrieval of interesting information that could not be considered exploiting a standard syntactic search. The experimental evaluation has shown that the presented approach performs better than traditional search methods.

In the future, further ways of exploiting semantics and user information in the search will be considered, for instance by adapting some of the ideas coming from past works in different contexts (e.g., semantic search in heterogeneous and dynamic graph data (Catania et al., 2013) and multimedia data (Grana et al., 2013)).

## Acknowledgements

This work was supported by the project “Algorithms and Models for Building context-dependent Information delivery Tools” (AMBIT) co-funded by Fondazione Cassa di Risparmio di Modena (SIME 2013.0660) and by the “SocialGQ” FAR 2016 FIM Department Project, University of Modena and Reggio Emilia.

## References

Abdou, S. and Savoy, J. (2008). Searching in medline: Query expansion and manual indexing evaluation. *Inf. Process. Manage.*, 44(2):781–789.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Beneventano, D., Bergamaschi, S., and Martoglia, R. (2016). Exploiting semantics for searching agricultural bibliographic data. *Journal of Information Science,* 42(6):748-762.

Bergamaschi, S., Martoglia, R., and Sorrentino, S. (2015). Exploiting semantics for filtering and searching knowledge in a software development context. *Knowledge and Information Systems*, 45(2):295–318.

Bolchini, C., Orsi, G., Quintarelli, E., Schreiber, F. A., and Tanca, L. (2011). Context modeling and context awareness: steps forward in the context-addict project. *Bulletin of the Technical Committee on Data Engineering*, 34:47–54.

Cabri, G., Gaddi, S., Martoglia, R. (2016). AMBIT-SE: Towards a User-aware Semantic Enterprise Search Engine. *In Proceedings of the 12th International Conference on Web Information Systems and Technologies (WEBIST 2016) - Volume 2*, pages 98–108. Springer.

Cabri, G., Leonardi, L., Mamei, M., and Zambonelli, F. (2003). Location-dependent Services for Mobile Users. *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems And Humans*, 33(6):667–681.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50.

Catania, B., Guerrini, G., Belussi, A., Mandreoli, F., Martoglia, R. and Penzo, W. (2013). Wearable Queries: Adapting Common Retrieval Needs to Data and Users (Vision Paper). *In Proceedings of the 7th International Workshop on Ranking in Databases (DBRank), 7:1-7:3.*

De Vocht, L., Softic, S., Verborgh, R., Mannens, E., Ebner, M. (2017). Social Semantic Search: A Case Study on Web 2.0 for Science. *International Journal on Semantic Web and Information Systems (IJSWIS),* 13(4).

Falcarin, P., Valla, M., Yu, J., Licciardi, C. A., Frà, C., and Lamorte, L. (2013). Context data management: An architectural framework for context-aware services. *Serv. Oriented Comput. Appl.*, 7(2):151–168.

Figueroa, A. and Neumann, G. (2016). Context-aware semantic classification of search queries for browsing community question–answering archives. *Knowledge-Based Systems*, 96:1-13.

Grana, C., Serra, G., Manfredi, M., Cucchiara, R., Martoglia, R. and Mandreoli, F. (2013). UNIMORE at ImageCLEF 2013: Scalable Concept Image Annotation. *In Proceedings of the Image Retrieval in Conference and Labs of the Evaluation Forum (ImageClef).*

Haslhofer, B., Martins, F., and Magalhães, J. a. (2013). Using skos vocabularies for improving web search. *In Proceedings of the 22nd International Conference on World Wide Web*, WWW ’13 Companion, pages 1253–1258.

Heflin, J. and Hendler, J. (2000). Searching the web with shoe. In *Artificial Intelligence for Web Search. Papers from the AAAI Workshop.*

Hyvonen, E., Saarela, S., and Viljanen, K. (2003). Ontogator: combining view- and ontology-based search with semantic browsing. In *Proceedings of XML Finland*.

Liu, F., Yu, C., and Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Trans. on Knowl. and Data Eng.*, 16(1):28–40.

Mangold, C. (2007). A survey and classification of semantic search approaches. In *Semantics and Ontology*.

Martoglia, R. (2011). Facilitate IT-Providing SMEs in Software Development: a Semantic Helper for Filtering and Searching Knowledge. In *Proceedings of the 23rd International Conference on Software Engineering and Knowledge Engineering* (*SEKE)*, pages 130–136.

Martoglia, R. (2015). Ambit: Semantic engine foundations for knowledge management in context-dependent applications. In *Proceedings of the 27th International Conference on Software Engineering and Knowledge Engineering* (*SEKE)*, pages 146–151.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communication of the ACM*, 38(11):39–41.

Ramona-Cristina, P., Vasilateanu, A., Goga, N. (2016). Ontology based multi-system for SME knowledge workers. *In Proceedings of the 2016 IEEE International Symposium on Systems Engineering (ISSE), Edinburgh, Scotland, October 3-5 2016,* pp. 1-5.

Rocha, C., Schwabe, D., and de Aragao, M. P. (2004). A hybrid approach for searching in the semantic web. In *WWW '04: Proceedings of the Thirteenth International Conference on World Wide Web*.

Savoy, J. (2005). Bibliographic database access using free-text and controlled vocabulary: An evaluation. *Inf. Process. Manage.*, 41(4):873–890.

Shekarpour, S., Marx, E., Ngonga Ngomo, A-C., Aue, S. (2015). SINA: Semantic interpretation of user queries for question answering on interlinked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30:39-51.

Thesprasith, O. and Jaruskulchai, C. (2014). Query expansion using medical subject headings terms in the biomedical documents. *In Intelligent Information and Database Systems - 6th Asian Conference, ACIIDS 2014, Bangkok, Thailand, April 7-9, 2014, Proceedings, Part I*, pages 93–102.

Villegas, N. M. and Müller, H. A. (2010). Managing dynamic context to optimize smart interactions and services. In Chignell, M., Cordy, J., Ng, J., and Yesha, Y., editors, *The Smart Internet*, volume 6400 of *Lecture Notes in Computer Science*, pages 289–318. Springer Berlin Heidelberg.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum*), pages 61–69.

Vu, T., Willis, A., Kruschwitz, U., and Song, D. (2017). Personalised Query Suggestion for Intranet Search with Temporal User Profiling. *In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 265-268

Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., and Li, H. (2010). Context-aware ranking in web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 451–458.

[i] AMBIT stands for "Algorithms and Models for Building context-dependent Information delivery Tools"

[ii] IPTC stands for "International Press Telecommunications Council", http://www.iptc.org/site/Home/

[iii] http://www.alfresco.com/

[iv] http://www.autonomy.com/

[v] http://lucene.apache.org/solr/

[vi] http://www.google.com/

[vii] http://www.attivio.com/

[viii] http://www.expertsystem.com/it/cogito/

[ix] http://www.coveo.com/

[x] https://www.ibm.com/

[xi] http://www.opensearchserver.com/

[xii] https://gate.ac.uk/

[xiii] http://www.cis.uni-muenchen.de/.17ex~schmid/tools/TreeTagger/

[xiv] http://wordnet.princeton.edu/ or, in case of specific contexts, other specialized resources

[xv] http://www.expertsystem.com/