Semantic Search in Portals using Ontologies

Wallace Anacleto Pinheiro
Ana Maria de C. Moura
Military Institute of Engineering - IME/RJ
Department of Computer Engineering - Rio de Janeiro - Brazil
[awallace,anamoura]@de9.ime.eb.br

Abstract.

Portals have become very popular on the internet. They allow users to access many online transactions and customized information. More recently, semantic portals emerged as a natural extension of traditional portals and they have been attracting researchers and enterprise attention as a new way to provide information. Ontologies are the basis of semantic portals, and one of their major functions is to categorize information inside portals. However, there are few practical applications showing effectively how and where ontologies can be used in this environment. This paper covers these issues and shows how ontologies can improve results relevance in one of the most important component of a portal: the search mechanism. It describes a search tool developed in the context of a semantic portal named PASS (Portal with Access to Semantic Search). This tool uses components of domain ontologies stored in the portal to expand terms during the search process. Its evaluation has been done using traditional unities of measure such as precision and recall, and average precision. Their corresponding evaluation graphics showed that searches performed according to our strategy returned more significant results than those generated by a traditional Web search mechanism.

1. Introduction

Portal is a recent concept that used to have a different meaning some years ago: it was considered simply as a search mechanism, whose purpose was to make information access on the internet easier. Later on some other functionalities have been included, changing these single search machines into complex environments, where users are able, from a single entry point, to have access to many different kinds of services such as: personalization, subject categorization, publishing and information distribution, a dynamic user collaborative environment, online transactions, etc.

Therefore, current traditional portals are neither able to share information with other portals, nor to present an efficient information retrieval strategy and metadata maintenance. Semantic portals arouse as an evolution of traditional portals [4, 8 12]. They aim at using ontology to provide more semantic expressivity to its information contents, as well as to improve some of its functionalities, as the search mechanism.

Users are usually disappointed when a large amount of irrelevant information is returned from a search on the Web. This fact may have different reasons, such as: documents and pages are not well described on the Web so that agents cannot index them properly; users do not contextualize his/her search, usually providing only one term to perform it.

This paper describes the semantic portal PASS, which is composed of two main modules: the ONTPASS, an ontology editor, and TOSS, a semantic search engine. The paper focuses on the TOSS engine, which includes a strategy to improve the search mechanism in the context of the semantic portal PASS. It uses domain ontologies to expand search arguments in order to increase results relevance. Terms expansion is done by transforming the original search and using concepts of the domain ontology. Evaluation showed that improvement of results relevance has been effectively achieved: search results performed by TOSS have been compared to those performed by a traditional search engine, such as Google, showing significant results.

The remainder of this paper is organized as follows. Section 2 gives a brief overview of semantic portals, and shows how ontologies can be inserted into these environments to improve search mechanism. Section 3 presents PASS, the proposed semantic portal environment, focusing on its semantic search component named TOSS. It describes the main approaches used to accomplish searches using ontology knowledge. Section 4 evaluates the search results performed by TOSS. It shows how the strategies employed in the search engine can improve results relevance. Section comments some related work, and finally, section 6 concludes the paper with additional comments and future work.

2. Semantic Portals and Ontologies

The attention of the computer science community has converged, due to the advent of the Web, towards the development of technologies that also enabled its use by machines, and not only focusing on humans. This new aim completely changed the way material should be published on the Web. It required standards and special mechanisms to represent semantic information content on the Web. Although still a challenge, researches progress towards the Semantic Web, such as envisioned by Berners-Lee [3].

Ontologies are central in the vision of the Semantic Web, since they enable knowledge sharing, reuse and common understanding between agents (human or machine), by providing a consensual and formal conceptualization of a given domain.

Semantic portals are evolving towards the Web Semantic generation [7, 12]. They aim at exploring semantics to provide and access information, as well as developing and maintaining their operational functionalities. Ontologies are fundamental to achieve these issues. In a portal environment, ontology can be seen as a taxonomy that defines terms and relationships among these terms, added of a set of rules, responsible for keeping some real semantic constraints. Hence, documents accessed either from a portal repository or from search engines on the Web are retrieved according to this ontology, taking into account their semantic contextualization.

The search engine in a portal is a very important component. Usually, current search engines take into account the hypertext analysis of the pages (links, structure, font size, etc.) to classify results relevance. In this context, some algorithms such as Pagerank [9], HITS [6] and some variations are used. The first algorithm considers a page as important whenever it reaches a certain number of page links that point at it. The latter includes, in addition to the page importance seen in [9], the pages they point at.

Therefore, the major problem of the existing search mechanisms concerns the large amount of irrelevant documents returned to the user. In fact, most of times this happens because users do not contextualize well the subject he/she wants to search, providing only one or two terms for the search engine. In this context, ontologies rise as a rich solution to help contextualizing user's search terms, since they make it possible associating concepts and properties around a specific domain.

3. A Portal with Access to Semantic Search

The portal PASS has been developed as an additional module built on the top of the KAON framework [4]. KAON is an open-source ontology management infrastructure, targeted for business applications. Its architecture, including the environment PASS, is shown in Figure 1.

KAON has been designed as a three-layered architecture (Client, Management and Storage), in order to incorporate new components (as plugins), ensuring modularity, extensibility and flexibility. Its main components are: OI-Modeler: an ontology graphical management tool that requires a complete installation in the user machine, and does not offer access from traditional browsers; KAON portal: enables creating Web portals based on ontologies; KAON API: is a set of interfaces to access KAON ontologies. Ontologies are represented in a RDF [11] based-format, through the RDF API component. Management and Storage layers are responsible for providing interfaces to access, manipulate and store ontologies in the system.

The environment PASS presents two major functionalities [10]: an ontology editor for Web named ONTOPASS (Ontologies for the PASS), and the semantic search tool named TOSS (TOol for Semantic Search). The first tool allows users to create and edit their own ontologies through a Web browser, an important item in a semantic portal. The KAON portal requires the ontology to be edited through its own specific interface, not through a Web browser.

TOSS uses terms of the ontologies stored in the portal to orient user's search.

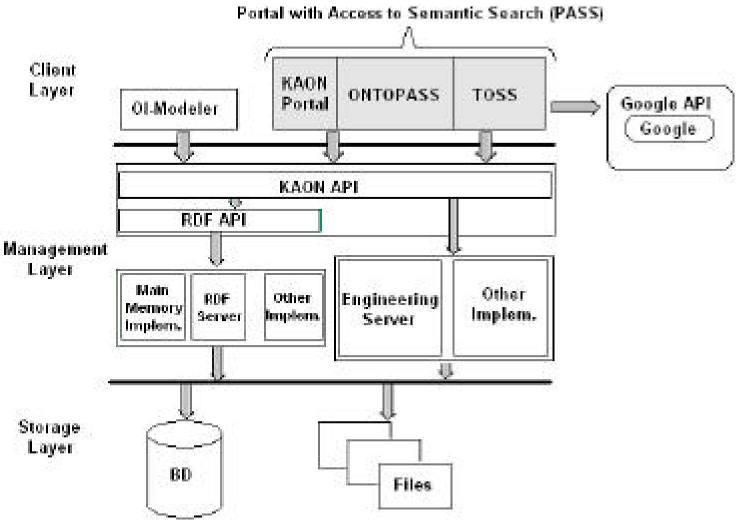


Figure 1 – Framework KAON with PASS

In fact the system creates a new sequence of terms, taking into account the original terms given by the user, and uses the Google[1] (through its API) to perform this new search. Its major purpose is to increase results relevance and to classify them according to the collaborative recommendation component.

Additionally this portal provides the ability: to store information about users and their preferences, allowing them to customize their pages. This enables performing queries on elements of that ontology domain, later and faster; and to achieve collaborative recommendation, where experienced users can recommend links to other users having less experience. PASS uses the KAON API in the management layer to access ontologies, which are stored in POSTGRES[2.]

The examples used in the context of this paper have their origin in a thesaurus of the Information Science domain[3]. In the literature the term thesaurus is usually employed as ontology. Although it is not considered as a full ontology, but as a light one [5], it can be very useful on helping to create ontologies, since it contains all concepts and relationships concerning a specific domain. Therefore, it is worthwhile to notice that a thesaurus does not provide inference rules, which are so important in ontologies.

3.1. The ONTOPASS Tool

One of the major advantages of this ontology editor concerns its utilization directly on the internet through a Web navigator, without requiring the installation of any other software. The whole ontology process edition occurs through navigation among concepts, properties and instances of the existing ontologies stored in the portal.

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

[2] http://www.postgresql.org/

[3] http://www.asis.org/Publications/Thesaurus/tnhome.htm.

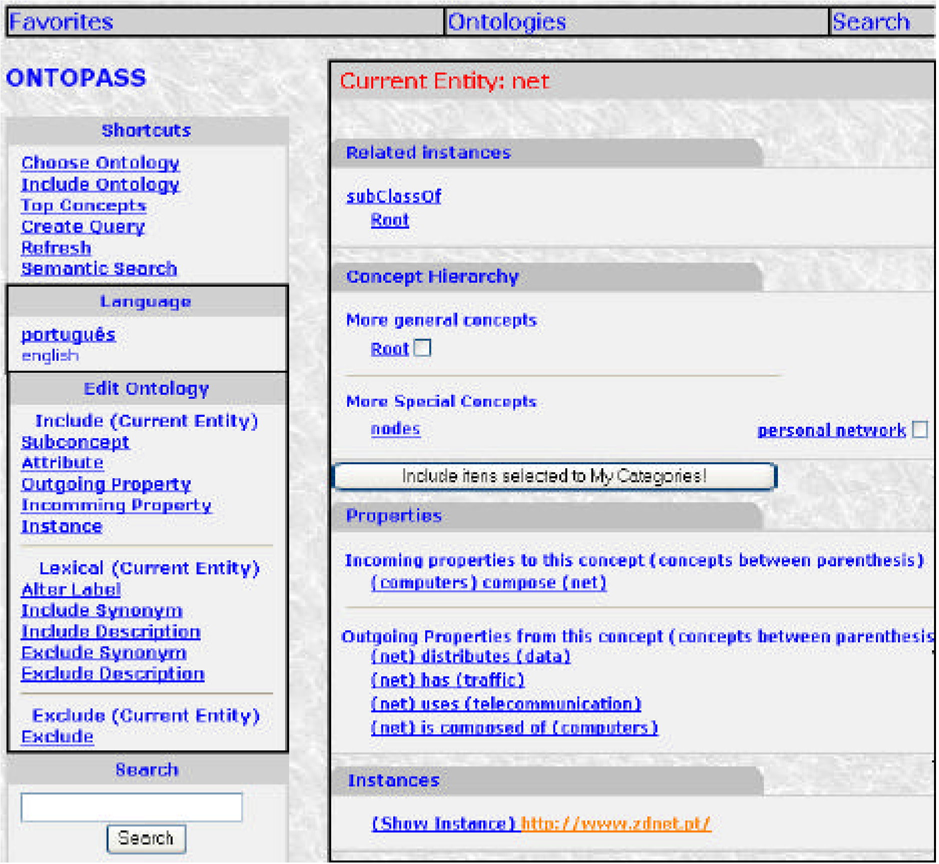


Figure 2 – ONTOPASS Interface

Relationships among concepts include different types of associations, such as hierarchical (generic/specific), aggregation (part-of) and generic (distributes, has, etc.), as well as synonyms, symmetric, transitive and inverse properties. Figure 2 presents the ontology interface for performing ontology edition.

The right side of the window shows some characteristics of the chosen entity, such as: super-concepts, sub-concepts, properties and instances, where it is possible for the user to customize some of the concepts as categories of his/her preference. At the left side, the user can choose the language he/she wants to interact with the system, as well as is able to edit the ontology concepts by including, removing and modifying their instances and properties.

3.2. The TOSS Tool

The main purpose of this tool is to offer a mechanism to help the user to contextualize his/her search. This is achieved by expanding the user’s search terms with other concept properties (subconcepts, synonyms and associated predicates) of domain ontologies stored in the portal.

For example, if a user wants to search the word “net” using current search agents, results will include links in the context of fishing, textiles, computers, among others. This ambiguity is caused by the lack of search context. From a portal a user can be introduced into a context, as his/her activities can be automatically analyzed, or information can be manually provided when he/she starts his/her search activities on the Web. This information is stored in the system and may be used afterwards to contextualize a search, and hence improve results precision.

Once in a user´s context, new terms can be associated to the set of terms originally chosen, providing a search that is enriched by the ontology semantics. For example, if the user is in the context of Computer

Science, when he/she asks for “net”, search will be performed taking into account other terms within this context, such as “computers” and “protocols”.

The main idea is to automatically relate these terms so that a new sequence of keywords is associated by the boolean operators AND/OR, also taking into account the concept properties and their semantic meaning. It is important to notice that current search machines consider a blank space as an implicit boolean operator AND. Symmetric, transitive and inverse properties are also used to order and associate search terms. However, when the term employed in the search is not a concept of the ontology, the search will be performed as in a traditional search, without using terms expansion.

The strategy used to expand terms in our tool takes into account the way concepts and properties are extracted from the ontology. It classifies them into three different groups of associations and could be considered a pre-processing query executed by the user [10], as described in the following sections.

3.2.1. Associated and Subordinated Concepts

This group is composed of the main concept provided by the user, the associated concepts (through any predicate) and specific concepts. As an example, suppose a user wants to search some information related to the topic “net”, and that the words “protocol” and “internet” have been defined as associated concepts (synonyms, for example) in the ontology. As “net” was the main term provided by the user during his/her search, the new sequence will always include it in the terms expansion process, added of the other associated concepts separated by the boolean operator “AND”. Synonyms will be considered in all groups of associations, and they will appear in the sequence separated by OR. Thus, the new resulting sequence, which is automatically generated from the terms expansion is:

net AND protocol OR internet

In fact, two set of sequences are searched separately:

net AND protocol

net AND internet.

It means that pages or documents containing both terms in each one of these sequences will be returned.

The use of AND after the term “net” imposes its presence and that of the following term (in the previous example “protocol” or “internet”) in the returned page. Its main purpose is to better contextualize the user´s term through the inclusion of new terms within the same domain. If the operator OR was used after the term “net”, the ambiguity problem of the original term would not be resolved, since new problems of ambiguity would be raised, even if terms were synonyms. For example, the term “protocol” can also refer to a set of rules concerning a person behavior, besides its meaning in the context of rules for data transfer. These two meanings would be associated if the operator AND was not included after the term “net”.

Additionally, suppose now that two other concepts are associated to “net”: “computers”, linked through the predicate “is composed of” and “traffic”, linked by the predicate “has”. Two sequence of terms are then generated:

net AND computers

net AND traffic

These sequences will be used to create a single sequence, which will also contain the synonyms showed before. Hence, the resulting term expansion is:

net AND protocol OR internet OR computers OR traffic

Now, consider that the concept “net” has sub-concepts defined in the ontology, corresponding to “nodes” and “personal network”. This fact offers two more possibilities:

net AND “nodes”

net AND “personal network”, resulting in the new sequence:

net AND protocol OR internet OR computers OR traffic OR nodes OR personal network

As a search machine considers the boolean AND as a blank space, the previous sequence becomes:

net protocol OR internet OR computers OR traffic OR nodes OR personal network

It is worthwhile noticing that the creation of this additional sequence provides a contextualization that is necessary to ensure relevance in the set of resulting pages expected from the user. In our example, as the term “net” is within the context of computers and networks, the search engine will discard all subjects related to fishing or textiles.

3.2.2. Predicates and Associated Concepts

This group is composed of the main concept and its associated concepts and predicates[1]. In order to exemplify this situation, consider figure 3, where the concept “net” is associated to “computers” and “traffic”, which are respectively connected to the predicates “is composed of” and “has”. From this figure it is possible to create the following sequences of terms:

net is composed of computers

net has traffic

Having considered predicates association, now assume the concepts “net” and “data”, which are associated to the transitive property “distributes”. Similarly, the concept “information” is linked to “data” through this same predicate, as shown in figure 4.

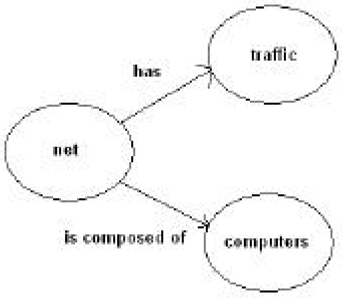


Figure 3 - An Oriented Graph with Associated Predicates

Taking into account the meaning of the transitive property “distributes”, it is possible to infer that if “net distributes data” and “data distributes information” then “net distributes information”, although these two concepts are not explicitly associated. Hence, two new sequences can be generated:

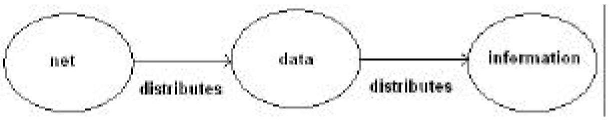


Figure 4 - An Oriented Graph with Associated Concepts and Transitive Property

net distributes data

net distributes information

Both sequences can be joined into a single one, as following:

net distributes data OR information

Once again the purpose here is to contextualize the search using existing properties and concepts of the ontology, requiring a minimum effort from the user.

3.2.3. Inverse Sequences

This last group is composed of the main concept, its associated concept and a symmetric or inverse predicate. Consider, for example, the predicate “is composed of” relating “net” and “computers” and having the inverse predicate “compose”. The following sequence of terms is generated for performing the search:

computers compose net, which represents the inverse of:

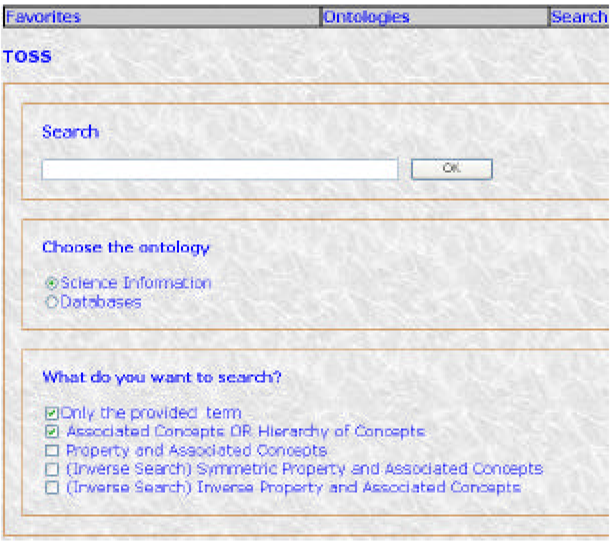
net is composed of computers

The symmetric property is also used in the expansion process. If the concept “net” is connected to “telecommunication” through the symmetric property “uses”, the following sequences are generated:

[1] In RDF a predicate is considered as a property

net uses telecommunication
telecommunication uses net

When applied to the ontology concepts, symmetric and transitive properties present interesting variations of search, although its use requires some especial attention: concepts found in the ontology turn out to have greater importance than the one chosen by the user, since they will occupy the first place in the sequence of terms that will be searched. This is one of the major points taken into account by the search engines, when classifying documents relevance before returning them to the user.



Choose the ontology

● Science Information
○ Databases

What do you want to search?

☑ Only the provided term
☑ Associated Concepts OR Hierarchy of Concepts
□ Property and Associated Concepts
□ (Inverse Search) Symmetric Property and Associated Concepts
□ (Inverse Search) Inverse Property and Associated Concepts

Figure 5 – Selecting Associations for Terms Expansion in TOSS

Figure 5 illustrates the use of terms expansion for the semantic search, where the user is able to choose which groups of associations he/she wants to consider in the search. These groups can be used simultaneously.

4. Evaluating the Search Proposal

Our strategy to evaluate results consisted of searching a set of concepts according to two procedures: initially using a tradition search engine, without considering any terms expansion process; then these results are compared to the results generated by TOSS, after submitting to it the terms expansion process described in section 3.

As already mentioned, the search tool Google has been chosen to make these experiments, as it provides an API that enables code adaptations and because it is currently considered as a very performing search engine [2]. Furthermore, as our tool performs the search process using three different groups of terms sequences, these groups have also been included in the comparative analysis of results. Twenty distinct concepts of the existing ontology have been previously selected to be submitted as search arguments. Additionally, only the first ten results provided by the engine have been taken into consideration for the experiments.

Measures such as precision and recall were used to evaluate results. These measure unities have been widely employed in Information Retrieval [1] to evaluate search engines. Figure 6 presents a graph containing the result curves corresponding to the search engines evaluated. Precision values have been calculated for each 3% of recall. That is, the higher the precision value for the same recall value, the better is the search engine in that recall point value. The graphic shows that the TOSS tool presents better results than those obtained by Google without terms expansion. Taking into account the strategies used by TOSS, in some points of the curve absolute gains have reached up to 15% and relative gains up to 102% (considering higher recall values) for precision values. The best result, considering the whole curve, concerns the approach that uses associated and subordinated concepts (3.2.1). The second best result is held by the approach using predicates and associated concepts (3.2.2). However, for inverse sequences curves did not show much significant difference when compared to results returned by Google.

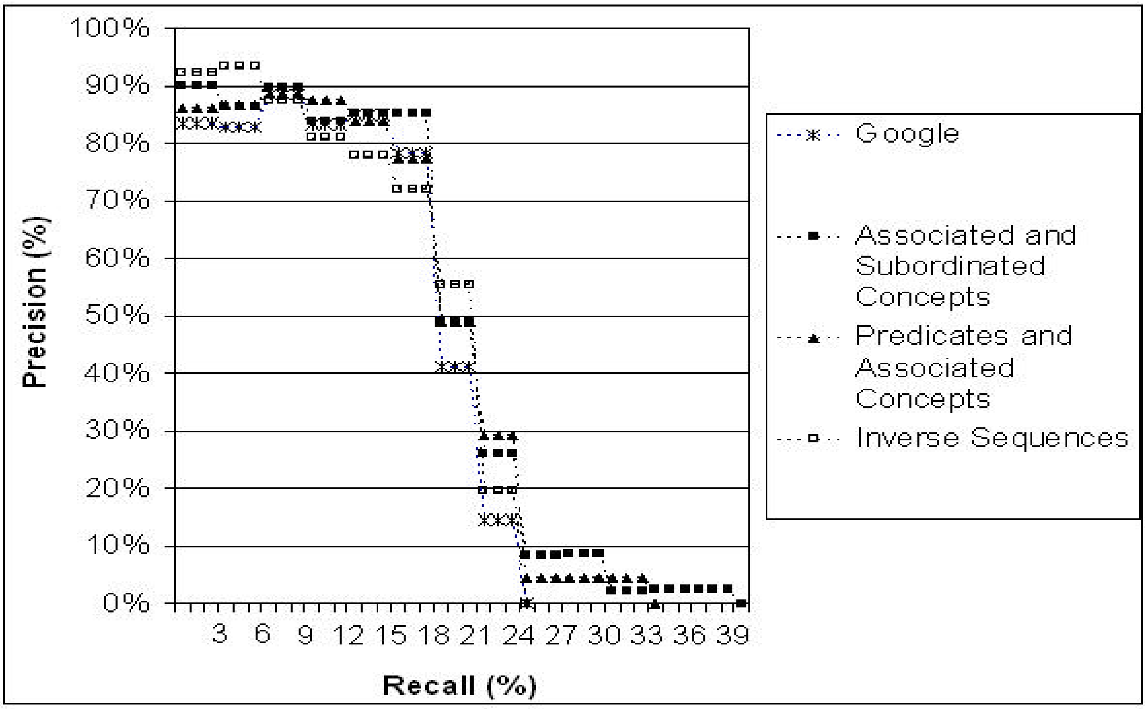


Figure 6 – Recall x Precision

Another type of measure, also used in the comparative analysis of search tools, was the average precision. It considers the average of the precision for the first result of all searches. Then it does the same for the second, and so on. For example, when we observe only the first result for each search, precision is 0% or 100%. If only two results for each search are observed, three different situations are possible: both results are relevant; only one of them is relevant or both are not relevant. Hence, for each search when only the first two searches are observed, precision is 0%, 50% or 100%. Similarly this process can be repeated to have the average precision for the first three results, four, and so on. These values can be plotted on a graph, where axis X presents the considered number of results given by the tool and axis Y shows the average precision percentage, as shown in Figure 7. This graph converges with the results presented in Figure 6. In some points of the curve TOSS had absolute gains of 15% and relative gains of 19% for average precision values. In all points of the curve, considering associated and subordinated concepts, the average precision of the engine TOSS presented higher or the same values as Google. It also presented higher results for predicates and associated concepts, and not significant difference for inverse sequences.

Generally speaking, it is possible to notice that the curve generated by predicates and associated concepts presents average precision results higher than the values provided by Google and by inverse sequences, holding the second place. Furthermore, we can observe that there is no dominant relation between the curve generated by Google and that generated by inverse sequences.

Taking into account associated and subordinated concepts, the number of relevant results produced by TOSS was 168, whereas in Google this number was 156, showing a gain of 5,75% in the total amount of relevant results.

However, when considering inverse sequences, the number of relevant results remained the same (156) for both Google and TOSS. This results can be partially explained, since the inversion between the main concept (the one that has been originally provided by the user) and the associated concepts during the terms expansion has much influence in the search: the latter sequence takes the first place in the terms sequence, assuming a higher priority for the search engine. Hence, tests showed that this inversion did not contribute much to improve the user's concept contextualization.

Despite this fact, an interesting fact was observed concerning the diversity of the pages returned to the user. None of the searches performed by the inverse sequence presented duplicate results when compared to the original search submitted to Google (without terms expansion). In the other cases the average of duplicate results was around 10%.

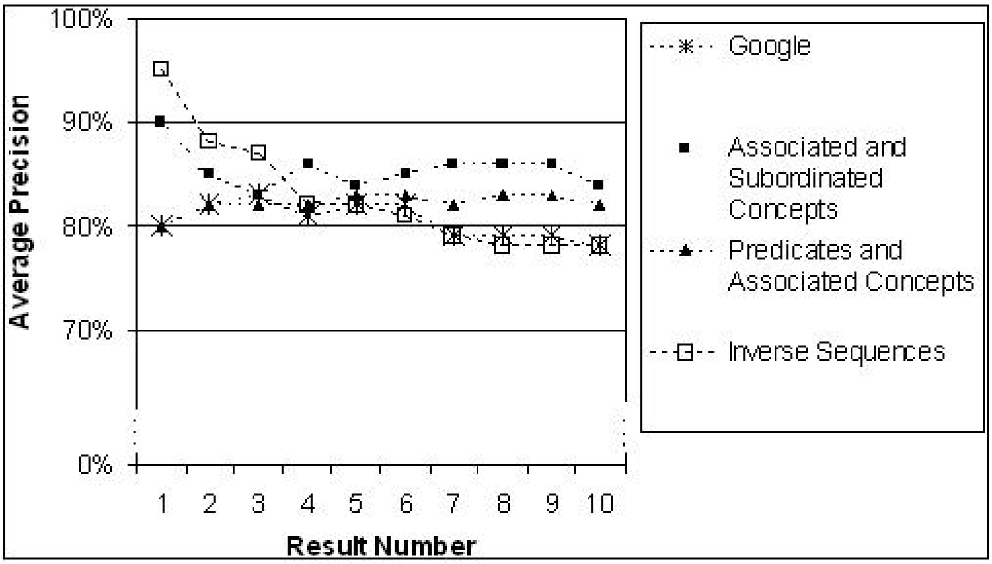


Figure 7 – Average Precision

5. Related Work

Some significant work has been developed on semantic portals. KA2 [12] offers a complete infrastructure providing resources, methodology and tools for ontology development especially designed to be used by heterogeneous research communities, geographically distributed.

The KA2 portal [12] is a special component of this environment. It provides an ontology view of the knowledge acquisition community. Besides of semantic retrieval, it allows comprehensive means for navigating and querying the knowledge base, and also includes guidelines for building such a knowledge portal. The potential users provide knowledge by annotating their Web pages in a decentralized manner. The knowledge is collected at the portal by crawling and presented to the user in different ways.

SEAL (SEmantic portAL) [8] is a generic approach for developing semantic portals. It exploits semantics for constructing and maintaining a portal, as well as providing information access at it. The knowledge base is represented in F-Logic.

The KAON Portal [4] is also an important initiative in the context of semantic portals, as described in section 3. It is a component of KAON framework and it acts as a simple tool for generating ontology-based Web portals. In order to create the portal, users need to create an ontology containing the information that will be presented in the Web. The KAON Portal may then be used to provide default visualization and navigation through this ontology.

Our work benefits from the KAON infrastructure, and it extends some of its functionalities in order to provide semantic search on the Web using a traditional and efficient search engine. In this context it exploits some other features that are found in SEAL, such as personalization views and establishes a collaborative recommendation mechanism, not found in these systems. Finally, another relevant contribution involved measuring the efficacy of using ontology to contextualize the arguments provided by the user during a search on the Web, increasing results relevance.

6. Conclusion

This work described the PASS environment, extended from the KAON portal, whose main purpose is to show how the use of ontologies can improve relevance of the results returned by a search engine on the Web. PASS is composed of two important modules: ONTOPASS, responsible for the ontology creation, edition and navigation; and TOSS, whose main function is to expand user's search terms according to a domain ontology stored in the system, to contextualize the search.

In order to show the efficacy of ontology use during the Web search process, we have done a comparative analysis of the results returned by the search engine using terms expansion from a domain ontology and not using them. This evaluation showed that when using approaches such as associated and subordinated concepts, as well as associated predicates, results relevance improves significantly, and hence its executions are worthwhile.

We are now working on the ONTOPASS tool in order to provide ontology import and a configuration strategy to enable its evolution. We also intend to improve the collaborative recommendation filter in PASS, in order to encourage communities of practice willing to share documents and relevant links in a certain domain on the Web.

References

1. R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval", Addison Wesley, New York, NY, USA, 1999.
2. J. Barker. "Google - the BEST Search Engine", UC Berkeley - Teaching Library Internet Workshops, 2003.
3. T. Berners-Lee, J. Hendler, O. Lassila. "The Semantic Web", in Scientific American, May 2001.
4. E. Bozsak, et. al. "KAON - Towards a Large Scale Semantic Web", Proc. of the 3rd Intl. Conf. on E-Commerce and Web Technologies (EC-Web 2002), 2002, pp. 304-313.
5. O. Corcho, et al. "Methodologies, Tools and Languages for Building Ontologies. Where is Their Meeting Point?", Data & Knowledge Engineering,46, 2003, pp. 41-64.
6. J. M. Kleinberg. "Authoritative Sources in a Hyperlinked Environment". Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms, 1998, pp. 668-677.
7. A. Maedche, S. Staab, N.Stojanovic, R. Studer, Y. Sure, "A Framework for Developing Semantic Web Portals", Lecture Notes in Computer Science, v. 2097, 2001.
8. A. Maedche, et al. "SEmantic portAL - The SEAL Approach", in Creating the Semantic Web, D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (eds.), MIT Press, MA, Cambridge, 2001.
9. L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Digital Libraries Working Paper, 1998.
10. W. A. Pinheiro. "Search in Semantic Portals: an Ontology-Based Approach" (in Portuguese), Master Thesis , IME, Feb. 2004.
11. RDF. "Resource Description Framework (RDF) Model and Syntax Specification", 1999, Available at: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.
12. S. Staab, et al. "Semantic Community Web Portals", 2000, in WWW9 / Computer Networks (Special Issue: WWW9 - Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, Maio, 15-19, 2000).