Praesto - A System for Contextual Semantic Search

Karina Fasolin[1], **Caio Stein D'Agostini**[1], **Renato Fileto**[1], **Renato Besen**[1]

[1]Departamento de Informática - Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brazil

{ kfasolin, csd, fileto, renatob}@inf.ufsc.br

Abstract. *This paper presents Praesto, a semantic search system that keeps track of the user's context as he poses keyword-based queries and browses the returned results. The context expresses the particular user's view and preferences for different denotations of keywords, which are described in an underlying ontology. It enables disambiguation and semantic extension of queries, in order to automatically generate search results that are semantically related to the user's preferences.*

1. Introduction

The availability of large quantities of information frequently overloads the user, who may not know how to manage or even understand the provided data [Fry 2007]. Furthermore, several information retrieval systems assume that the users' needs are static, being unable to adapt to changes of interests, which are natural for people [Baeza-Yates and Ribeiro-Neto 1999]. Thus, the users are not always able to benefit from all the available information, in part because of poor machine-human communication.

Traditional information retrieval systems are based on lexical matching of user-provided keywords with their occurrences in the indexed contents and/or markup [Baeza-Yates and Ribeiro-Neto 1999]. These systems do not recognize the intended meaning of user-provided keywords. Thus, they frequently return results related to keyword denotations (meanings) that the user is not interested in. In addition, semantically relevant results may be omitted if they are not lexically related to the provided keywords.

These problems, which affect precision and recall, can be alleviated by using an ontology to describe the possible meanings of keywords, and trying to identify the user's intentions in a semantic search system [Mangold 2007]. However, catching the exact user's intention is a communication intensive task. The use of suitable communication and visualization techniques in the human-computer interface can help the user to understand the body of knowledge related to a particular domain or application and better express his intentions [Michlmayr et al. 2007].

This paper presents Praesto, a semantic search system designed to capture and exploit users' context information, relative to an ontology, in order to provide semantically customized and more relevant results for individual users. It keeps track of the evolution of the user's context by capturing information from the interactions of the user with the system when posing queries and browsing results.

Praesto maintains the context of each user as a partial view of the knowledge described in an underlying ontology. This context is customized according to the users preferences for specific denotations of the keywords used in previous searches. By presenting

the users knowledge view collected in his context and classifying the search results accordingly in the machine-human interface, Praesto helps the user to understand alternative meanings of keywords and better communicate his intentions. Also, as more contextual information is collected, the system has better conditions to automatically solve ambiguities and semantically extend queries.

The remainder of this paper is organized as follows. Section 2 presents Praestro's general architecture and provides an overview on how it manages context information. Section 3 describes Praestro's Web-based user interface and how it presents the users' contexts and the search results. Section 4 discusses some related work. Section 5 closes the paper with some conclusions and the enumeration of future works.

2. Contextual Semantic Search

Context in information retrieval systems can be understood as the relevance of each subject for the user and how subjects are related from the user's viewpoint [D'Agostini et al. 2008]. The user's context represents information capable of affecting how the system interprets requests from that user.

The system presented in this paper is part of a project aimed to improve results of semantic searches by using contextual information to complement formal descriptions provided by a domain ontology [D'Agostini et al. 2008, D'Agostini and Fileto 2009]. Praestro organizes information in three layers, as illustrated by the triangle on the left of Figure 1: (i) the user's context, which is different for each individual user; (ii) the shared ontology, which provides formal definitions for the alternative meanings of terms in a particular domain to be worked with; and (iii) the contents to be searched for, which are semantically annotated with terms defined in the ontology.

Praestro maintains each user's context in a topic graph $TG = G(T, E)$ [D'Agostini et al. 2008]. Each topic $t(weight) \in T$ refers to a particular denotation of a keyword, described in the underlying ontology, and each edge $e(t_1, t_2, weight) \in E$ represents a semantic relation between topics. The $weight$ of a topic $t(weight)$ represents the degree of the user's interest for t. The $weight$ of an edge $e(t_1, t_2, weight)$ is a measure of how much the topics t_1 and t_2 are semantically related from the user's viewpoint.

The topic graph (TG) is created by observing the user's interactions with the system for posing queries and browsing results. The topics corresponding to keywords that return relevant results have their weights increased, as well as the associations between topics referring to ontology terms used to annotate a result item. This means that, if the user selects irrelevant results and ignore the relevant ones, Praestro receives wrong information for updating the context, which can affect the quality of future searches.

Figure 1 presents an overview of the Praesto's searching process, where the tasks involving user interactions are represented by ellipses. The user specifies a query in the *Search* task, by providing one or more keywords. Two or more keywords in the same query are considered as disjunctive (i.e., connected with OR). For each user provided keyword, the system first verifies if this keyword is present in the users' TG. If so, the system employs the TG to drive the semantic search according to the user's preferences.

If the searched keyword is not found in the TG, the system looks for that keyword in the underlying ontology. If ambiguous definitions are found in the ontology, the sys-

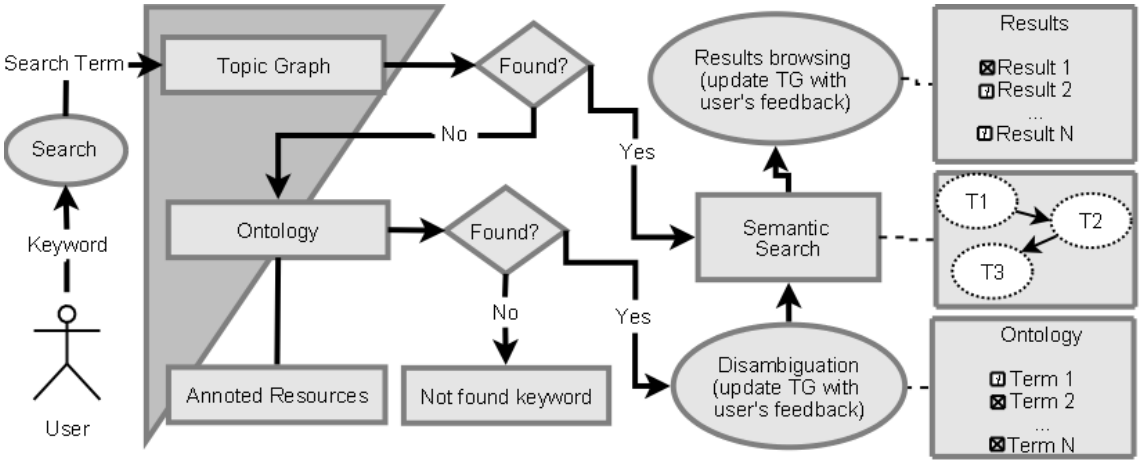

Figure 1. Praestro searching process.

tem inquires the user through the *Disambiguation* task, asking for the denotation(s) he is interested in. A list of options helps the user to differentiate alternative denotations and to choose the one(s) which are in accordance with his intentions in the particular search. The denotations chosen by the user are inserted in the TG, in order to be used in semantic searches. When a keyword is not found in the ontology, it is stored in a repository of unknown terms, for future evolution of the ontology.

After processing all the keywords from a query, the system uses the possibly updated TG to drive the recovering process, classify and rank the results according to the user's preferences. In the *Results browsing* task, the user can mark the results that interest him. It gives feedback for the system to update the TG accordingly. Then, the user can restart the process from the *Search* task again, by posing the same query or a new query for the system to process with the updated TG. More details of the contextual semantic search process and about the evolution of the context graph as the user interacts with the system are presented in [D'Agostini and Fileto 2009].

3. Praestro Graphical Interface

This section presents how a proper presentation of the keyword denotations, the user's context and the search results can help the user to understand the provided information, and enable the system to capture high quality feedback, and consequently obtain better results. The visualization problem, in this case, can be divided into two sub-problems [Michlmayr et al. 2007]: the first is centered in the user profile (profile-centric), and the former in data (data-centric). In order to better illustrate these issues, first consider the following example.

Suppose that a user is searching for the keyword *São Paulo*. This query is ambiguous, as *São Paulo* can refer to a *city*, a *state*, or a *soccer team*, among other meanings. Furthermore, each denotation of *São Paulo* can have synonyms or be closely associated with other subjects. The denotations to be considered in the query depend on the user's intentions towards the keyword.

3.1. Profile-Centric: Disambiguation and User's Context Visualization

When the TG does not contain any topic whose label matches the searched keyword, and the system finds more than one denotation for the keyword in the underlying ontology, the user is questioned about his intentions. He is asked to choose one or more of the alternative denotations of the keyword as illustrated in Figure 2 (A).

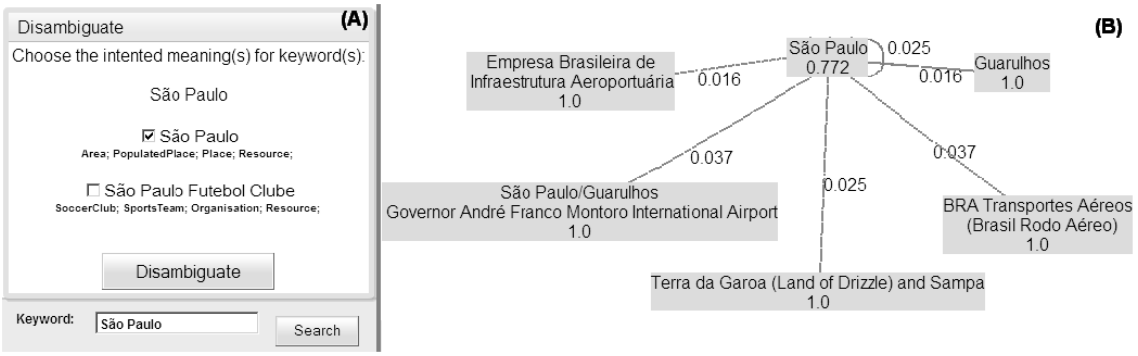

Figure 2. (A) The disambiguation dialog ; (B) The weighted topic graph of a user.

Suppose that the user chooses the first denotation of *São Paulo*, because he is interested in *São Paulo's Airport*, instead of *soccer*. The system can use this denotation to update the TG, and then perform the search. Figure 2 (B) shows the user's TG after some interactions with the system searching for an airport close to the city of *São Paulo*. For simplicity, some topics are hidden in this context. The context information in the graph indicates that the results related to the *airport* that serves the *city* of *São Paulo* should be given a higher relevance than other results. That is indicated by the weights of the topics. The higher the weight, more likely is the interest of user for the corresponding topic or association.

Context visualization helps the user to understand ambiguous denotations of keywords, and relations among denotations. It also provides an idea of how searches are processed, why each result is retrieved and how results are ranked, according to the information the TG. It can also lead to means for the users to materialize and communicate their understanding of a given domain.

3.2. Data-Centric: Search Results Visualization

Figure 3 presents the results returned by Praestro to answer the search for *São Paulo* when the user's TG is as presented in Figure 2 (B). The user can select the results that look interesting for him and search again, causing the context to change accordingly. Then, the system process the search again with an updated topic graph.

The presentation of the search results plays an important role for the quality of the context information management, and consequently for the quality of the searches. If too many results are presented without a clear organization, the user will most likely not look through all the results with the same level of attention, as users dislike dealing with long disorganized lists of results [Baeza-Yates and Ribeiro-Neto 1999].That is an undesired situation, since it results in low quality feedback from the user to the system.

In order to help the users identifying and understanding the search results, the system organizes the results according to two criteria: (i) single ranking defined by the

relevances present in the user's context, as shown in the left of Figure 3; and (ii) multiple rankings, each one for a different denotations of a keyword, as shown in the right of Figure 3. Through these two different visualization options the user can either see the results ranked according to their relevance to the search or he can separate the results according to specific subjects of the ontology.

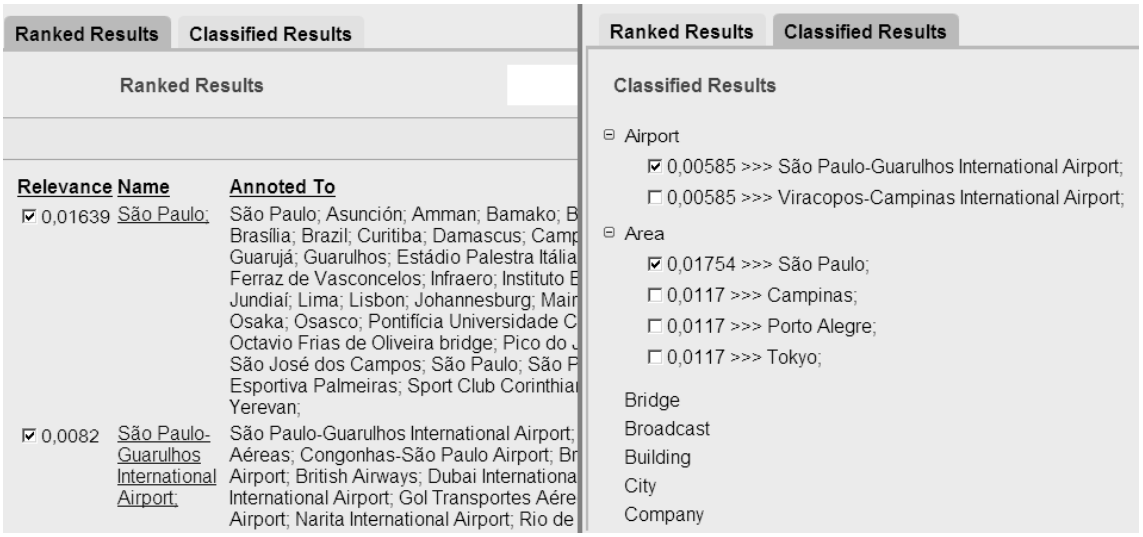

Figure 3. Results organized by general rank (left) and denotations (right)

4. Related Work

Several systems have been proposed recently to support semantic search on the Web, such as Freebase[1] , Hakia[2], Cuil[3], and Powerset[4]. These systems take into account the possible meanings of the provided keywords to process the queries and present the results classified accordingly. They use slightly different approaches for semantically processing queries, ranking and presenting results, and allowing the user to disambiguate keywords and refine their queries. Praestro has the distinguished characteristic of collecting context information about each individual user's preferences, with respect to keywords denotations described in an underlying ontology, in order to automatically solve ambiguities and provide customized results for the user, besides rendering a graphical visualization of the user's context.

The visualization of the user's context helps the users to comprehend semantic relationships among denotations of provided keywords. This understanding can be useful for query refinement and assimilation of classified and ranked results. The use of graphical visualization for information retrieval systems is supported by the ability that humans have to easily assimilate graphs, instead of long text lines. [Koshman 2006] provides an analysis of an information retrieval system that uses graphical interfaces to present results. This interface is developed under the hypothesis that information visualization systems are attractive to humans because they draw upon the efficiency of the human visual processing abilities.

[1]http://www.freebase.com
[2]http://www.hakia.com
[3]http://www.cuil.com
[4]http://www.powerset.com

5. Conclusion and Future Work

Praestro is a contextual semantic search system with the following functionalities: (i) keeping track of context information for each individual user in a weighted topic graph (TG); (ii) using the TG to disambiguate and semantically expand queries, in order to return customized results for each user; (iii) collecting feedback from the users interactions with the system, in order to keep the topic graph aligned with the evolving users intentions; (iv) allowing the user to visualize his TG, to be aware of semantic relations between denotations of keywords, and to better communicate his intentions to the system.

The TG hides details of the potentially huge and cumbersome ontology used to annotate the resources. It allows each user to access his individual view of the knowledge present in the ontology, narrowing the gap between the users' knowledge and the shared knowledge formalized in the ontology.

Further information about Praestro can be found in http://www.lisa.ufsc.br/Praestro. Future works on this system include: (i) allowing the user to specify searches by selecting topics and associations in his TG; and (ii) allowing the user to define topics referring to concepts that were not originally described in the ontology.

6. ACKNOWLEDGMENTS

This work is supported by CNPq (grant 48139212007-6), Fapesc and FEESC. Special thanks to members of Grupo-DB/UFSC and LaPeSD/UFSC for their valuable help.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

D'Agostini, C. S. and Fileto, R. (2009). Capturing users' preferences and intentions in a semantic search system. *The 21st International Conference on Software Engineering and Knowledge Engineering - SEKE*.

D'Agostini, C. S., Fileto, R., Dantas, M. A. R., and Gauthier, F. A. O. (2008). Contextual semantic search - capturing, using the user's context to direct semantic search. In Cordeiro, J. and Filipe, J., editors, *ICEIS (4)*, pages 154–159.

Fry, O. (2007). *Visualizing Data*. O'Realy Media Inc.

Koshman, S. (2006). Visualization-based information retrieval on the web. *Library and Information Science Research*, 28(2):192–207.

Mangold, C. (2007). A survey and classification of semantic search approaches. *Int. J. Metadata, Semantics and Ontology*, 2:23–24.

Michlmayr, E., Cayzer, S., and Shabajee, P. (2007). Tech report: Hpl-2007-72: Adaptive user profiles for enterprise information access. Technical report, HP Labs Technical Reports.