An Interactive Graph Based Legal Information Retrieval System

Paulo Quaresma and Irene Pimenta Rodrigues

Departamento de Informática,
Universidade de Évora,
7000 Évora, Portugal
{pq|ipr}@di.uevora.pt

Abstract. A graph based information retrieval framework for the visualization of clusters of related legal documents is proposed. Clusters are obtained through the analysis of the citations and topics of the retrieved documents.
Users, interactively, may refine their queries by selecting the graph nodes. Then, the context of the users interaction is calculated and it is shown in a tree-structure. Users are allowed to return to previous interaction states by selecting these tree nodes.
This framework was applied to the decisions of the Portuguese Attorney General and it was made available in the web (http://www.pgr.pt – in Portuguese [QR99]). A detailed example of a user interaction with the system is shown in the paper.
An evaluation procedure was applied to the system and it showed a decrease of the average number of interactions per user session.

1 Introduction

Legal information retrieval systems are increasing their complexity, namely, the size and number of their text bases. Moreover, there is a special need for IR systems that are able to calculate and to show the relations between the different legal texts. In fact, legal texts can be related via different aspects: citations, topics, and author.

In this paper we present a web legal information retrieval system that is able to cooperatively calculate and show relations between the retrieved set of documents. These relations are shown using graphs and the users are allowed to refine their queries by selecting the graph nodes.

Cooperation is also achieved through the visualization of the interaction context in a tree-structure. Users are allowed to return to previous interaction states by selecting some of these tree nodes.

This framework was applied to the decisions of the Portuguese Attorney General and it was made available in the web (http://www.pgr.pt – in Portuguese [QR99]). The framework was implemented in a logic programming environment that has been used previously with success to model rational agents and their interactions [APP+00].

Section 2 briefly describes the core information retrieval module. Section 3 describes the clustering procedures that were used and section 4 describes the construction of the context structure. In section 5 a detailed example is presented and in section 6 some evaluation results are presented. Finally, in section 7 conclusions and future work are pointed out.

2 Information Retrieval Module

The legal information retrieval system is based on SINO [GMK97] from the AustLII Institute. SINO was changed in order to be adapted to the Portuguese Language. Namely, the new system uses the Portuguese lexicon (more than 900,000 words) in order to handle morphological errors and to obtain the base queried word. For instance, if the user asks to be informed about documents where a specific word appear, the systems also searches for documents containing derived words (plurals for nouns, verbal forms for verbs, ...).

As a top layer over the basic IR system we are using a juridical terms thesaurus. This thesaurus is a result from another project: PGR - Selective access of documents from the Portuguese Attorney General.

The juridical terms thesaurus can be described as a taxonomy which has the relations:

- is equivalent to
  ex: law is equivalent to norm
- is generalised by
  ex: prime minister is generalised by minister
- is specified by
  ex: accident is specified by traffic accident
- is related with
  ex: desertion is related with traffic accident

The thesaurus is used to expand queries to include all the values that are equivalent or more specific or related, with the initial query (more information can be found in [QR99]).

The result is a powerful IR system, which has many similarities with the work described in [BvWM+99], namely, allowing the extraction of textual information using localization, inference, and controlled vocabulary.

3 Document Clustering

In our framework we decided to cluster documents based on two different characteristics: citations and subjects. All documents were previously analyzed and a database relating each document with its citations and subjects was built. Then, for each user query, a set of documents is obtained (using SINO) and these documents are clustered using the relations previously calculated. Finally, the obtained clusters are visualized as a list or as a graph structure (using the

Graphviz package developed by AT&T and Lucent Bell Labs). For a complete description of the cluster process, namely its methodology and algorithms, see [QR00].

3.1 Citations

The complete set of legal documents from the Portuguese Attorney General was processed to obtain the citations between documents.

In order to obtain all the citations it was necessary to construct a specialized chart parser, which is able to partially analyze the text and to retrieve and to normalize the citations between documents. Note that documents can be cited by their number, date, title, author, and by almost any mixture of these fields. Taking this fact into account, we have built a database with these fields for each document and, whenever a possible citation appears in the text, the parser checks if the cited document exists and it builds a new citation entry in the database.

These citations are used by the system to build the graph of relations between the set of documents retrieved by the user queries.

As an example, the query "bombeiro" (fireman) obtains the following graph of citations:

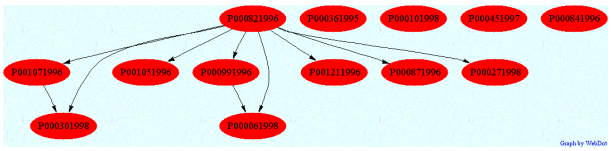

Fig. 1. Citations: Bombeiro – Fireman

Note that it was possible to detect an important legal document that is referred by most of the other retrieved texts. It was, probably, a document that created jurisprudence about some fireman cases. The user is able to directly search that document by "clicking" in the document node.

3.2 Topics

In order to obtain the clusters of topic relations it was necessary to classify each document accordingly with a set of concepts previously defined by the Portuguese Attorney General Office (PAG Office). The classification was done manually by the PAG Office and automatically by a neural network [QR00]. The documents were parsed in order to build the topics relationship database. Using this database it is possible to visualize the graph of topic relations and/or a list of clustered concepts.

Graph of topic relations The graph of relations is calculated using the database of topic relations that was manually and automatically previously built. Each pair of documents with, at least, 90% of common topics are related by a graph arc.

As an example, the query "bombeiro" obtains the following graph of topic relations:

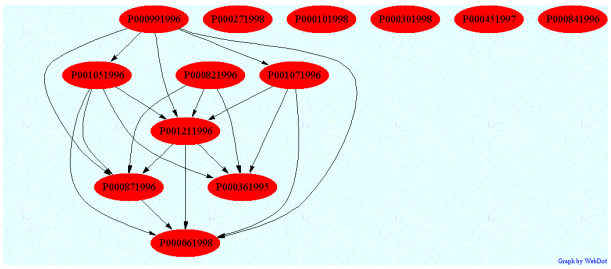

Fig. 2. Topics: Bombeiro – Fireman

As in the previous section it was possible to detect a cluster of closely related documents and a set of non related documents (probably about some minor distinct subjects). As it can be seen, using this approach users are cooperatively helped in their searches.

4 Context Structure

The user interaction context is kept in a tree structure. This structure records both user and system questions and answers. The structure is used to compute the meaning of a user query and to allow the user to return to a previous point of the dialogue and to build a new branch from there.

As an example of the use of the interaction structure suppose that, after the query "bombeiro", the user wants to refine the query by the concept "ferido" (hurt). Figure 3 shows this interaction structure (in Portuguese).

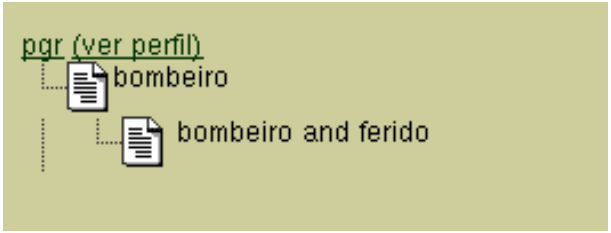

Fig. 3. Query: Bombeiro – Fireman; Refine by "ferido – hurt"

The visualization of the tree-context structure allows the user to easily select non-explored branches of the tree and to refine his previous queries.

In the example, if the user queries the system with the expression "militar" (military), the system may be able to detect that, as there are no documents about "fireman", "hurt", and "military", the intended user query may be the refinement of a previous query: "fireman", and "military". Figure 4 shows the tree-structure after this inference.

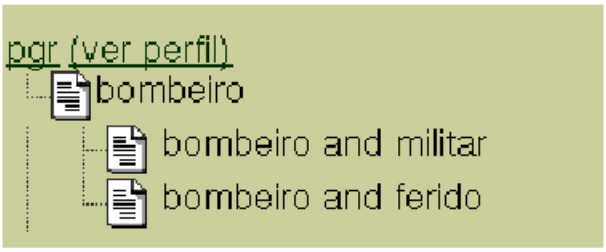

Fig. 4. Query: Bombeiro – Fireman; Refine by "militar – military"

In the next section a more detailed example is presented.

5 Interaction Example

Suppose the user wants to find documents about militaries that have been injured and received a pension from the government. He wants to find out if there have been similar cases in the legal knowledge base of the Portuguese Attorney General.

In order to help readers and to keep the example short, the Portuguese web version is not presented. Instead an English compacted version of the interaction is shown. The complete web interaction can be easily reproduced from the URL http://www.pgr.pt

[User - Q1:] Militares (militaries)

[System - Q2:] Result: 1127 documents

Documents with keyword:
(120) incapacity
(86) medal
(28) court
(16) commission
(11) rights
(4) marriage
...

Options on this list were obtained by clustering the set of 1127 documents.

[User - Q3:] Chooses: incapacity

[System - Q4:] Result: 120 documents

Documents with keyword:
(10) service
(6) pension
...

Options in this column are obtained by re-clustering the 120 documents.
[User - Q5:] Chooses: pension
[System - Q6:] Result: 6 documents

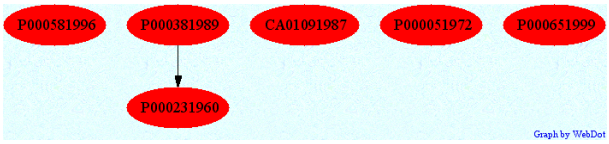

Fig. 5. Graph clustering by topic

Options in figure 5 were obtained by graph clustering by topic the 6 documents.

[User - Q7:] Refinement: illness (after analyzing the topics of the retrieved documents and deciding that they were not relevant for his goals)

[System - Q8:] There are no documents of the 6 above that match your new constraint, I assume that you want your query to refine Q3, i.e. you are looking for documents about: militaries and incapacity.

[User - Q9:] Result: 6 documents

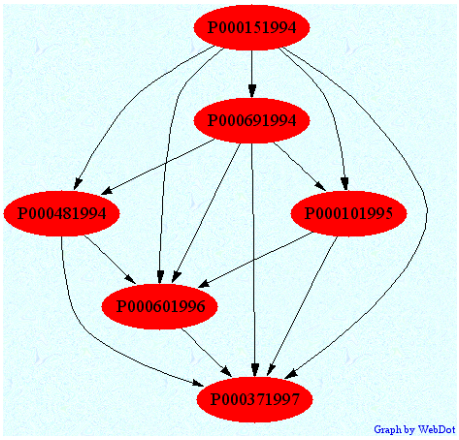

Fig. 6. Graph clustering by topic

Options in figure 6 were obtained by graph clustering by topics the 6 documents.

.....

This example shows some of the flexibility and cooperativeness of the system, allowing users to dynamically refine their queries and helping them in a proactive way, giving hints and clustering the retrieved documents.

During the interaction, the tree representation of the dialogue is being inferred and displayed in a visual tree-diagram. As an example, the tree representation of the previous example is presented in figure 7 (in a compact version).

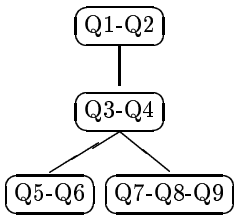

Fig. 7. Interaction Structure Tree

6 Evaluation

As it is widely known, the evaluation of legal systems with knowledge representation and reasoning capabilities are a very complex and difficult task [SZ99].

The main goal of this evaluation process was not to evaluate the information retrieval module but the impact of the graph visualization of clusters in the interaction with the users.

We have defined two sets of users: one using the clustering module and the other accessing directly the information retrieval module. We have recorded their queries during the second semester of the year 2000 and the preliminary results show that our system is able to help the users decreasing the average number of queries needed to obtain the desired documents (around 20%).

7 Conclusions

We have presented a legal information retrieval system that uses graphs to visualize clusters of related documents. The visualization of these clusters may help the users in their searches, showing the relations between the documents and clustering them by topic.

Moreover, the interaction context is inferred and displayed in a tree-structure. This visualization allows the users to analyze the history of the interaction and to explore other branches of the structure.

The preliminary evaluation results, showed that our cooperative system is able to help the users decreasing the average number of queries needed to obtain the desired documents (around 20%).

As future work, we intend to obtain more evaluation results (quantitative and qualitative) of the cooperative graphical clustering module. We also would like to apply our system to other information retrieval systems, namely to non-Portuguese legal documents.

References

[APP+00] J. J. Alferes, L. M. Pereira, H. Przymusinska, T. C. Przymusinski, and P. Quaresma. Dynamic knowledge representation and its applications. In S. Cerri and D. Dochev, editors, *Proceedings of the 9th International Conference on Artificial Intelligence - Methodology, Systems, Applications (AIMSA'2000)*, number 1904 in Lecture Notes in Artificial Intelligence, pages 1–10, Varna, Bulgary, September 2000. Springer Verlag.

[BvWM+99] Tania Bueno, Christiane von Wangenheim, Eduardo Mattos, Hugo Hoeschl, and Ricardo Barcia. Jurisconsulto: Retrieval in jurisprudencial text bases using juridical terminology. In *Proceedings of the ICAIL'99 – 7th International Conference on Artificial Intelligence and Law*, pages 147–155. ACM, June 1999.

[GMK97] G. Greenleaf, A. Mowbray, and G. King. Law on the net via austlii - 14 m hypertext links can't be right? In *In Information Online and On Disk'97 Conference, Sydney*, 1997.

[QR99] P. Quaresma and I. Rodrigues. Pgr: A cooperative legal ir system on the web. In Graham Greenleaf and Andrew Mowbray, editors, *2nd AustLII Conference on Law and Internet*, Sydney, Australia, 1999. Invited paper.

[QR00] Paulo Quaresma and Irene Pimenta Rodrigues. Cooperative Information Retrieval Dialogues through Clustering. In Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *Proceedings of the Third International Workshop on Text, Speech and Dialogue—TSD 2000*, Lecture Notes in Artificial Intelligence LNCS/LNAI 1902, pages 415–420, Brno, Czech Republic, Sep 2000. Springer-Verlag.

[SZ99] Andrew Stranieri and John Zeleznikow. The evaluation of legal knowledge base systems. In *Proceedings of the ICAIL'99 – 7th International Conference on Artificial Intelligence and Law*, pages 18–24. ACM, June 1999.