Cooperative Information Retrieval Dialogues Through Clustering

Paulo Quaresma and Irene Pimenta Rodrigues

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

Abstract. In this paper we present some aspects of a cooperative web information retrieval system in the law domain. Our system is able to infer the user intentions and to keep the context of the user interaction in order to supply suggestions for further refinement of the user query. One important aspect of our system is its ability to compute clusters of documents associating a keyword to each cluster. A detailed example of an interaction with the system is presented.

1 Introduction

In this paper we present some aspects of a cooperative web information retrieval system with juridical documents based on SINO, a boolean text search engine from the AustLII Institute [GML97].

During an interaction when a user poses a query we want that our system will be able:

- To infer what are the user intentions with the queries [Loc98, QR98, QL95, Pol90].
  When a user asks for documents with a particular keyword, usually he is interested in documents that may not have that keyword and he is not interested in all documents with that keyword.
- To supply pertinent answers or questions as a reply to a user question.
  The system must supply some information on the set of documents selected by the user query in order to help the user in the refinement of his query.

In order to accomplish this goals we need:

- To record the previous user interaction with the system (user questions and the system answers).
  This record will play the role of a dialogue structure [CL99, RL93]. It provides the context of sentences (questions and answers) [CCC98], allowing the system to solve some discourse phenomena such as anaphoras and ellipses. Since our system is multi-modal, other user acts such as button clicks and menu choices are also represented in our dialogue structure.

- To obtain new partitions (clusters labelled with a topical keyword) of the set of documents that the user selected with his query(ies).
- To use domain knowledge whenever the system has it.

In our system each event (utterance) is represented by logic programming facts that are used to dynamically update the previous model. Using this approach it is possible to represent new events as logic programs and to obtain the new states. Moreover it is possible to reason about past events and to represent non-monotonic behaviour rules.

Each utterance will trigger the inference of the user intentions taking into account the user attitudes (such as his beliefs and the user profile). The results of the inference of the user intentions are:

- A new set of user and system beliefs and intentions (such as the intention of the user to be informed of something by the system)
- A new dialogue structure. This structure keeps the dialogue context allowing for the interpretation of user acts in its occurrence context. The dialogue structure constraints the interpretation of user intentions and it is built as a result of the intentions inference.

2 Information Retrieval System

During an interaction the user wants to look for some documents and his queries are ways of selecting sets of documents. The system questions and answers always intends to help the user in his search of documents by supplying information on subsets of documents in the text database.

After a user query the system may:

- Show the set of documents selected by the query.
- Present a set of keywords that may help the user to refine his query. In order to build a set of keywords the system may build groups of documents (clusters) from the initial set selected by the user query (with or without expanding).
- Present a set of concepts that may help the user to refine his query. In cases where the system as knowledge about some of the documents subject it is possible to build groups of documents using that knowledge, and to provide the user concepts for refining its query.
- Explain the user why his query does not select any document, providing suggestions for other queries.

2.1 An Interaction Example

Dialogue 1 Suppose that the user his a retired fireman that has saved a child from drowning in a river during his holidays putting his life at risk. Since he has caught a pneumonia as a consequence, he wants to know if he can get a pension for relevant services to his country.

[User - Q1:] Documents about pensions for relevant services?

[System - Q2:] There are 243 documents, listed in the right frame. For further refinement you may choose one (or none) option in each column:

Pensions that were:	*Pensions for:*
given or refused	militaries or civilians

Documents with keyword:
drowning or medal or in duty or ...

Options in column 1 and 2 are obtain using rules that model legal knowledge on legislation on pensions for relevant services. Options on column 3 are obtained by clustering the set of 243, using the method described in section 4.

[User - Q3:] chooses: civilians and drowning.

[System - Q4:] There are 45 documents, listed in the right frame. For further refinement you may choose one (or none) option in each column:

Documents with keyword:
death or accident or ...

Options in this column are obtain by reclustering the 45 documents.

[User - Q5:] Documents where there was a drowning in a river.

[System - Q6:] Since there are 13 documents of the 45 above that match the concept, I've listed them in the usual frame. For further refinement you may choose one (or none) option in each column:

Documents with keyword:
car crash or bridge or ...

[User - Q7:] Where there was someone swimming?

[System - Q8:] Since there are no documents of the 13 above that match your new constraint. I assume that you want that your query refines Q3, i.e. you are looking for documents with the concept swimming, and drowning and pensions for relevant services for civilians. Do you want to see these documents?

[User - Q9:] Yes. .....

3 Interaction Context

The interaction structure supplies the context for the user and system actions. This representation structure takes into account that an utterance may: specify the information contained in a set of previous utterances; or to open a new context, when the user does not intend to continue refining its query and desires to start a new one.

The Interaction structure (IS) is made of segments that group sets of acts (user and system sentences). The Interaction structure reflects the user intentions; it is built taking into account the user and system intentions. The Interaction segments have precise inheritance rules defining how segments heritage their attributes from the attributes of their multimodal actions.

An example of the interrogation context for dialogue 1 is shown in fig. 1.

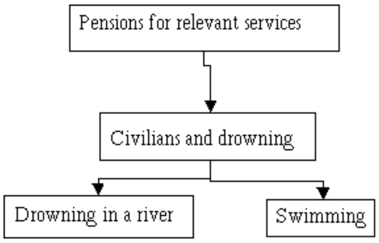

Fig. 1. Interrogation context after utterance Q7

4 Intelligent Clustering

Clustering is a complex process [Sal89] since it involves: the choice of a representation for the documents, a function for associating documents (measures for similarity of documents with the query or between them) and a method with an algorithm to build the clusters. One of the best clustering methods is the Scatter/Gather browsing paradigm [CDRKT92, CKP93, HP96] that clusters documents into topically-coherent groups. It is able to present descriptive textual summaries that are built with topical terms that characterise the clusters. The clustering and reclustering can be done on-the-fly, so that different topics are seen depending on the subcollection clustered.

4.1 Clustering and reclustering

Given a set of documents selected by a user query, a structure associating a set of descriptors to each document (the document classification) is built, structure 1, with a linear (O(n), n is the number of texts) procedure. This structure is transformed in another structure, structure 2, that associates to each descriptor in the first structure a set of documents, with a procedure that has complexity O(n*m), m is the number of descriptors in the structure. These structures are shown in fig. 2.

Finally we must must choose a set of descriptors that:

1. The union of the set of documents associated to the descriptors is the initial set of documents.

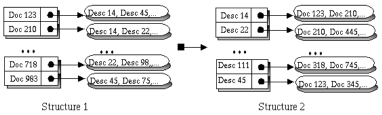

Fig. 2. Document structure

2. The intersection of the set of documents associated to any two descriptors is empty.

These two conditions can not be satisfied always, when this is the case, the first one is dropped.

However there are other proprieties the set of descriptors should have:

1. Its cardinality should be between 10 and 20.
2. The cardinality of each document set should be similar.
3. Descriptors with only one document associated should be ignored.

Our search space, for m descriptors in structure 2, will have 2m states that should be tested. Since it is not possible to search all the state space in a reasonable time we have to use some heuristics in order to cut off part of the search space, and we use an informed search algorithm, a best first search with an evaluation function specially designed for this problem.

This procedure will start by:

– Sort structure 2 by descendent order of the cardinality of the documents set.
– To eliminate the descriptors with only one document associated.
– To represent each set of documents in a bit table, to simplify the test for inclusion of document (it will became O(1)).

Then the best first search will be guided by an evaluation function that always choose to add a descriptor that as a set of documents with its cardinal as near as possible of the interval [10, 20].

The search ends with success when:

– All documents are selected, the union of the sets associated with the selected descriptors is the set of selected documents.
– The cardinal of the set of descriptors reaches 30, and the cardinal of the union of the sets of documents is greater then 70% of the initial number of documents.

Evaluations of this algorithm grants that it will take O(n*m), n is the number of documents, m is the number of descriptors in structure 2 without those eliminated in the first step. For 10000 documents and 2000 descriptors it will take 100 milliseconds, a reasonable time for a search in World Wide Web.

The reclustering can be done by modifying structure 2, taking out the documents that are not selected in the refinement, and resorting this structure. Normally reclustering is must faster the initial clustering, since the input is smaller, and structure 2 is already there.

5 Conclusions and Future work

We aim to build a cooperative IR system for juridical information. In order to be cooperative, to help the user to find a specific set of documents, our system needs to represent some knowledge about the database documents. One important source of knowledge is obtained by clustering sets of documents and labelling each cluster with a topical term resulting from a document juridical analysis.

By now the evaluation of our system has been performed by a set of users, mainly law students, that think that the systems suggestions are helpful for their searches. We hope to use other evaluations criteria that may quantify how helpful can the system suggestions be, but by know we only have a quality evaluation.

References

[CCC98] J. Chu-Carroll and S. Carberry. Response generation in planning dialogues. *Computational Linguistics*, 24(3), 1998.

[CDRKT92] D. R. Cutting, J. O. Pedersen D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proc. 15th Annual Int'l ACM SIGIR Conf. on R&D in IR*, June 1992.

[CKP93] D. R. Cutting, D. Karger, and J. Pedersen. Constant interaction-time scatter/gathern browsing of very large document collections. In *Proc. of the 16th Annual Int. ACM/SIGIR Conf.*, Pittsburgh, PA, 1993.

[CL99] Sandra Carberry and Lynn Lambert. A process model for recognizing communicative acts and modeling negotiation subdialogs. *Computational Linguistics*, 25(1), 1999.

[GML97] G. Greenleaf, A. Mowbray, and G. King. Law. On the net via austlii 14m hypertext links cant be right? In *In Information Online and on Disk97 Confrence*. Sydney, January, 1997.

[HP96] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis:scatter/gather on retrieval results. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference*, Zurich, June 1996.

[Loc98] Karen E. Lochbaum. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4), 1998.

[Pol90] Martha Pollack. Plans as complex mental attitudes. In Philip Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communications*. MIT Press Cambridge, 1990.

[QL95] P. Quaresma and J. G. Lopes. Unified logic programming approach to the abduction of plans and intentions in information-seeking dialogues. *Journal of Logic Programming*, 54, 1995.

[QR98] P. Quaresma and I. P. Rodrigues. Keeping context in web interfaces to legal text databases. In *Proc. of the 2nd French-American Conf. on AI&LAW*, Nice, France, 1998.

[RL93] I. P. Rodrigues and J. G. Lopes. Building the text temporal structure. In *Progress in Artificial Intelligence: 6th EPIA*. Springer-Verlag, 1993.

[Sal89] Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Addison-Wesley, 1989. Reading, MA.