Automatic Classification and Intelligent Clustering for WWWeb Information Retrieval Systems

Paulo Quaresma and Irene P. Rodrigues
Departamento de Informatica
Universidade de Evora
7000 Evora, Portugal
(pq|ipr)@di.uevora.pt

Abstract

In this paper we present some aspects of an intelligent interface for a WWWeb legal information retrieval system.
Our system is able to keep the context of the user interaction in order to supply suggestions for further refinement of the user queries.
The set of documents obtained from the user queries is dynamically organised in clusters labeled with keywords from a juridical thesaurus. Since, some of the texts were not previously classified, we have developed an automatic juridical classifier based on a neural network. The classifier receives as input a legal text and proposes a set of juridical terms that characterize it.

Keywords: Legal web−based Information Retrieval System, clustering of documents, automatic classification

1. Introduction

In this paper we present some aspects of an intelligent interface for a WWWeb information retrieval system with juridical documents in more then one text database.
Our information retrieval system is based on SINO, a boolean text search engine from the AustLII Institute [GMK97]. The text databases are built with the Portuguese Attorney General documents, Supreme Court Decisions and other Court instances Decisions.
The structure of our texts are similar (several fields such as: number, date, administrative information, conclusions, full text, etc.) except for the field with the juridical analysis. The texts from the Portuguese Attorney General are classified using the set of keywords from the juridical thesaurus. The other documents do not have that field, so our system has to automatically build it. This classification of texts is obtained using a neural network that has been trained with the texts from the Portuguese Attorney General that were already classified.
We have developed an automatic juridical classifier based on a neural network. The classifier receives as input a legal text and proposes a set of juridical terms that characterize it. The proposed terms belong to a taxonomy of juridical concepts developed by the Portuguese Attorney General Office.

During an database interrogation 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,LA87]. It provides the context of sentences (questions and answers) [CCC98,Wal96], 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 labeled 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
- a new set of user and system 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 is built as a result of the intentions inference.

In order to model the knowledge about the documents in the texts database the system represents four levels of knowledge using dynamic logic programming. The knowledge levels are: Interaction, Domain, Information Retrieval and Text.

The interaction level is responsible for the dialogue management. This includes the ability of the system to infer user intentions and attitudes and the ability to represent the dialogue sentences in a dialogue structure in order to obtain the semantic representation of the dialogue.

The domain level includes knowledge about the text domain and it has rules encoding that knowledge. For instance, in the law field it is necessary to represent under which conditions a pension for relevant services may be given to someone; those pensions are usually attributed to militaries or to civilians such as firemen, doctors, and nurses.

The Information Retrieval Level includes knowledge about what we should expect to find in texts about a subject, for instance that in texts about pensions for relevant services, the pension may be attributed or refused. This knowledge is obtain trough the clustering of the documents selected by a query. Our process of clustering and reclustering requires that all documents are classified with a set of keywords (expressions) of our juridical thesaurus (see section 4). This classification can be obtain using a neural network as it is explained in section 5.

The Text Level has knowledge about the words and sequence of words that are in each text of the knowledge base.

2. Information Retrieval System

In our system the user wants to look for some documents and his queries are ways of selecting sets of documents. The system questions and answers always intend 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.
Since our information retrieval system is boolean, the documents that are selected are just those that match the query and there is no particular ordering in the way they are listed.

Show the set of documents selected by the expanded query.
Our information retrieval system has options for expanding a query such as:
- expand using morphologic flexion: verbs, nouns, adjectives, etc.;

- expand using synonyms (a general dictionary);
- expand using a domain thesaurus.

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 expansion).
These groups of documents may be selected by keywords (expressions) that define them.
In order to supply as much information as possible to the user, the groups of documents are disjoint, i.e., there are no document that belongs to more then one group.

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.
Most information retrieval system assume that the user will never get to a dead−end with its queries, so they relax on the meaning of a query by using non boolean retrieval systems.

These systems are appropriate for general Web searches, but for a system that intends to control the search we think that a boolean system is best suited.
A user query must be interpreted in context. It is through the inference of user intentions that our system interprets the user query in its context.
The system takes into account user profiles in the inference of user intentions. Since our database has juridical documents, we have different profiles for layers, law students and other users.

Dialogue 1
Example of a dialogue with a user that has profile *other users*. 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:	Documents with keyword:
Given or refused	militaries or civilians	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

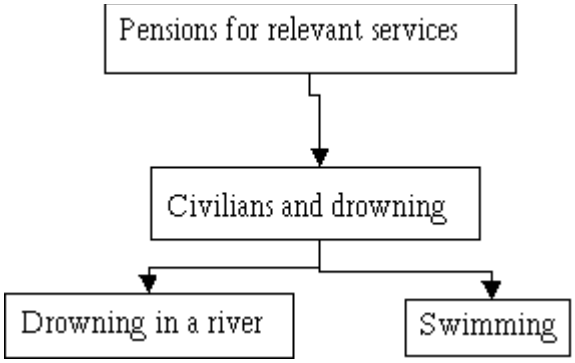
The Interaction context will use a knowledge level, which has rules for the multimodal interaction. It includes:

1) the rules for inferring the user intentions necessary to generate the system acts;
2) the rules necessary to build the interaction structure in order to obtain the representation of the user acts and its context.

Cooperation with the user is achieved through the inference of user intentions. The system tries to infer the user intentions in order help him to find out the set of documents that the user is looking for.
The system helps the user by informing him about the domain knowledge (juridical) and particularities of the texts in the knowledge base. This way the user is guided by the system in the task of refining his queries.
The interaction representation 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.
The Interaction structure is built by recognizing the user intentions and using them in order to enable the system to intervene in the dialogue using pertinent multimodal acts.

Example of the interrogation context for dialogue 1, after utterance Q7 :

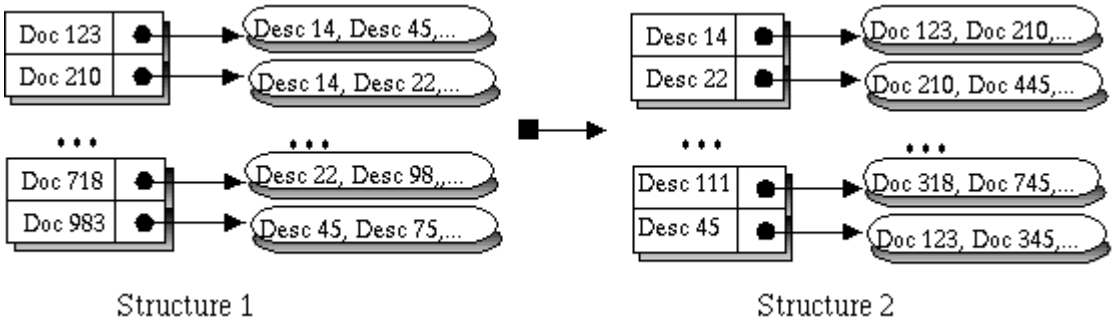


4. Intelligent Clustering

A service that an information retrieval system always provide is the organization of retrieval results. Most of the systems rank the output according to estimated relevance values. But when there are large document groups with similar rankings some systems try to build groups (clusters) and to label the groups with some relevant keywords.
Clustering helps users by showing them some kind of pattern in the distribution, it allows the user to include or exclude sets of documents from further searches.
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 build with topical terms that characterize the clusters. The clustering and reclustering can be done on−the−fly, so that different topics are seen depending on the subcollection clustered.
We use the topical terms describing groups of documents in our natural language interface as knowledge used to inform the user on possible further choices for defining its final goal.
The result of clustering the output of the information retrieval system is used in our reasoning processes as a knowledge base containing knowledge about the current documents in the juridical text database. The topical terms that characterize the clusters are chosen from the expressions in our juridical thesaurus.

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.



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.
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 that the set of chosen descriptors should have, namely:

1. Its cardinal should be between 10 and 20, but it must be always greater then 1 and less then 30.
2. The cardinal of each document set associated to each descriptor in this set should be similar.
3. Descriptors that only have one document associated with it should be ignored.

Our search space, for m descriptors in structure 2, will have 2^m 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 last entries, those that have 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 the near as possible of [number_of_documents/20, number_of_documents/10].

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 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 does 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. Automatic Classification

As it was seen in the previous section it is very important that all documents are classified accordingly with a juridical taxonomy. In fact this classification is the basis of the intelligent clustering and the system query refinement suggestions.

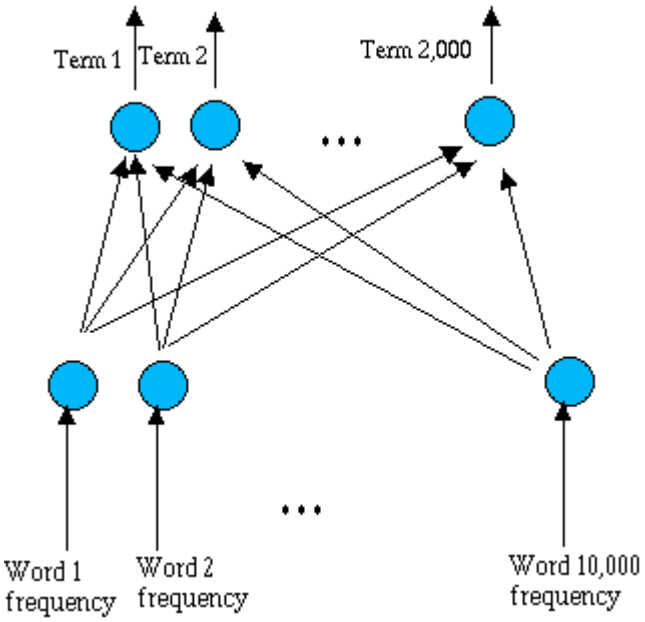
As described previously, our information retrieval system is based on a legal database composed by texts from the Portuguese Attorney General and from other legal sources (Supreme Court Decisions and other Court Instances).
However, some of the texts do not have a juridical analysis field, i.e., they were not previously classified accordingly with a juridical taxonomy. In order to handle this situation, we have developed an automatic juridical classifier based on a neural network. The classifier receives as input a legal text and proposes a set of juridical terms that characterize it. The proposed terms belong to a taxonomy of juridical concepts developed by the Portuguese Attorney General Office.

The classifier was developed using the Stuttgart Neural Network Simulator [SNNS98] and the network is a feed−forward network with two layers of units fully connected between them. The first layer has the input units and the second layer has the output units. Each input unit is associated with a specific word and its input value is the word frequency in the text. Each output unit is associated with a juridical term and its value is 1 or 0, defining if the juridical term characterizes the input text.

In order to build the network it was necessary to create a mapping between the text words and the input units. After analyzing the legal texts we obtained a set of 10,000 distinct words, composed only by nouns, verbs, adverbs, and adjectives. In this process we have discarded all the other word classes and we have reduced each word to its canonical form (infinitive for verbs, singular for nouns). We

have also mapped each used juridical term (2,000) to a specific output unit. Finally, connections between all input units and all output units were created: 10,000 x 2,000 = 20,000,000.

The following figure shows the network topology:



As learning algorithm for this feed−forward neural network, it was used the standard backpropagation algorithm and the net was trained until the average squared error per unit per pattern was less than 0.025.
The training set was composed by 95% of the texts from the Portuguese Attorney General and the validation set was composed by the other 5%. We also have a test set composed by other legal texts not previously classified.
As results for the validation set we obtained that 90% of the proposed terms were correct. It is possible that the other terms are not completely incorrect. In fact they may be a different characterization of the text and they have to be analyzed by juridical experts.
Regarding the test set with other legal texts we only have some preliminary results, which point to a worst performance of the network. The evaluation of the proposed terms by our partners of Portuguese Attorney General suggests that on average each document has 80% of the terms correctly associated, 20% are wrongly associated, and there are 10% of terms missing. This behavior can be explained by noticing that the network was only trained with texts from a specific source (Attorney General). However, our results are very promising, and validates the proposed classification tool (the neural network) as a very useful tool for automatic classification of juridical texts for an information retrieval system.

3. Conclusions

We aim to build an intelligent interface for juridical information retrieval systems. 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 obtain by clustering sets of documents and labeling 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.

Since we want to expand our system with other juridical databases that do not have a juridical analysis using the terms of our juridical thesaurus we needed to build one tool for automatic

documents classification.
We have built an automatic classifier for Portuguese legal texts based on a feed−forward neural network with 12,000 units and 20,000,000 connections. This network was able to classify documents obtaining 90% of correct answers. However, much work needs to be done: a) training the net with texts from different sources and with different domains; b) evaluate different topologies and different learning algorithms.

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.

[GMK97] G. Greenleaf, A. Mowbray, G. King. Law on the Net via AustLII 14M hypertext links cant be right? In Information Online and on Disk97 Conference, 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.

[LA87] Diane Litman and James Allen. A plan recognition model for subdialogues in conversation. Cognitive Science, 11(1), 1987.

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

[MH95] S. McRoy and G. Hirst. The repair of speech act misunderstandings by abductive inference. Computational Linguistics, 21(4), 1995.

[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.

[SNNS98] SNNS − Stuttgart Neural Network Simulator, version 4.2, User Manual. University of Stuttgart, 1998.

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

[Wal96] Marilyn Walker. The effect of resources limits and task complexity on collaborative

planning dialogue. Artificial Intelligence, 85(1), 1996.