Personalizing information retrieval in CRISs with Fuzzy Sets and Rough Sets

Germán Hurtado Martín°*, Chris Cornelis*, Helga Naessens°
° Hogeschool Gent
* Ghent University

Summary
Current Research Information Systems (CRISs) usually contain large amounts of heterogeneous and distributed data, which makes finding specific information difficult for a user. It is in these cases that the concept of a personal search agent, proactively informing the user about newly available information, becomes more and more popular. But how can the agent know what is useful for the user if he has not expressed it explicitly? Our approach proposes using fuzzy and rough sets to make the matching process between the users' interests and the information in the system more flexible, as they allow expressing partial relationships and expanding queries, as well as dealing with problems like imprecision, ambiguity, or incompleteness.

1 Introduction

Due to the growing information volume in CRISs, complicated by the fact that typically these systems are updated on a daily basis, users are often not aware of new information relevant to them. As a consequence, they are often also unable to express their information need accurately by means of a traditional query. In order to prevent a user (e.g., a researcher) from missing out on useful information (e.g., opportunities to apply for research funding), a personal search agent can be used to support the search process and to present meaningful recommendations to the users. Applications of this idea can already be found in research mobility portals such as ERACareers (EraCareers 1995), or, outside the field of research, the idea of this kind of agents has become very popular among job sites, which share the above-mentioned problems (large information volume, frequent updates) with CRISs.

In general, an agent tries to match a user to information items (documents), based on his profile and those of other users. Usually the matching process looks in the documents for exact matches of terms of interest (keywords) in the users' profiles. This approach is subject to limitations as often the documents use different terms to refer to the same, or similar, concepts. This can be mitigated by mapping documents to a predefined term collection (a taxonomy, or thesaurus), with the caveat that this operation is likely to incur a certain loss of precision. Furthermore, most CRISs also face other information defects such as missing, or ambiguous, information.

To tackle the above problems, we propose a more flexible matching process, governed by concepts from fuzzy (Zadeh 1965) and rough (Pawlak 1982) set theory. Fuzzy sets allow to express partial relationships (between terms, between users, as well as between users and taxonomy terms), which adhere closer to reality than a binary (black-and-white) classification. The rough component provides mechanisms for query expansion: in this way, a user profile and a document may still be matched when they refer to related keywords, resulting in a higher recall, and — provided the connections between taxonomy terms are chosen sensibly— no significant loss in precision. Furthermore, both fuzzy and rough sets perform well in the presence of imperfect or missing information, and their hybridization provides a fruitful query refinement mechanism (De Cock et al. 2007; Magalhães et al. 2006).

At Hogeschool Gent, a Personal Alert System (PAS) is being developed along these lines. PAS combines profiles of the staff and their research activities, with documents about funding possibilities to support research. Both these information sources are mapped to a common ontology, which is currently a version of the three-level IWETO taxonomy (Iweto 2007), enriched with a fourth level containing free keywords obtained from staff profiles. The main goal of the system is to alert users whenever a funding opportunity can be matched to their research interests by the search agent. A basic prototype has been implemented, deploying the algorithms in (Srinivasan et al. 2001), and in which the user can also influence term relations through a simple feedback process.

The structure of this paper is as follows: in Section 2, we recall important concepts about fuzzy sets, rough sets, and fuzzy rough sets. In Section 3, we present PAS, giving an overview of its architecture; we mainly focus on the mapping facility, the central part of the system, and also illustrate how it operates. In Section 4, we address some issues for future work. Finally, the paper is concluded in Section 5.

2 Theoretical background

2.1 Fuzzy Sets

Unlike classical set theory, where objects belong to a given set or not, fuzzy set theory (Zadeh 1965) allows that objects belong to a set or relation to a given degree. A fuzzy set in X is defined as an $X \rightarrow [0, 1]$ mapping, while a fuzzy relation in X is a fuzzy set in $X \times X$. For all y in X, the R-foreset of y is the fuzzy set Ry defined by $Ry(x) = R(x, y)$ for all x in X. If R is a reflexive and symmetric fuzzy relation, that is, $R(x, x) = 1$ and $R(x, y) = R(y, x)$ hold for all x and y in X, then R is called a fuzzy tolerance relation. For a fuzzy tolerance relation R, Ry is sometimes called the "fuzzy similarity class" of y.

The intersection of fuzzy sets is usually modeled by the minimum operation, that is, for fuzzy sets A and B in X, we define $A \cap B$ by $(A \cap B)(x) = min(A(x),B(x))$. The bounded difference of A and B is defined by $(A \,|\!-\!|\, B)(x) = max(0, A(x) - B(x))$. Finally, the α-cut of A can be defined as the classical set $A_\alpha = \{x \in X \mid A(x) \geq \alpha\}$, where $\alpha \in [0, 1]$. It contains those elements which belong "sufficiently" to the fuzzy set A.

2.2 Rough sets

Rough sets (Pawlak, 1982) provide approximations of concepts in the presence of incomplete information. In particular, rough set analysis makes statements about the membership of some object y of a universe X to the concept of which A is a set of examples, based on the indiscernibility between y and the elements of A. Usually, indiscernibility is modeled by means of an equivalence relation R on X. Its equivalence classes $[x]_R$ can be used to define the lower and upper approximation of A by

$$R{\downarrow}A = \{x \in X \mid [x]_R \subseteq A\}$$
$$R{\uparrow}A = \{x \in X \mid [x]_R \cap A \neq \emptyset\}$$

Hence, the lower approximation is the union of all equivalence classes contained in A, while the upper approximation is the set of equivalence classes which have a non-empty intersection with A. This means that the lower approximation is the set of objects that can be positively classified as members of A, while the upper approximation is the set of objects possibly belonging to A. The couple $(R{\downarrow}A, R{\uparrow}A)$ is called a *rough set* in X.

2.3 Fuzzy rough sets

Research on the hybridization of fuzzy sets and rough sets emerged in the late 1980's, and has focused mainly on fuzzifying the formulas for lower and upper approximation seen in the previous section. To this aim, the set A is generalized to a fuzzy set in X, allowing that objects can belong to a concept to varying degrees. On the other hand, rather than assessing objects' indiscernibility, we may measure their closeness, represented by a fuzzy relation R. Typically it is assumed that R is at least a fuzzy tolerance relation.

The lower approximation $R{\downarrow}A$ and the upper approximation $R{\uparrow}A$ of A by means of R are defined in this paper by

$$(R{\downarrow}A)(y) = \inf_{x \in X} \max\,(1\text{-}R\,(x, y), A(x))$$
$$(R{\uparrow}A)(y) = \sup_{x \in X} \min\,(R(x, y), A(x))$$

for all y in X. More general definitions have been proposed; we refer e.g. to (Radzikowska & Kerre, 2002).

Fuzzy rough sets have applications in machine learning, but are also useful in other contexts such as query expansion (Srinivasan et al., 2001, De Cock & Cornelis 2005) for improved information retrieval. This is connected to some of the problems discussed in the introduction, and a first application of fuzzy rough sets in the context of CRISs was presented in (Magalhães et al. 2006).

3 Architecture of PAS

As mentioned in the introduction, the Personal Alert System (PAS) at Hogeschool Gent stores information about projects and funding opportunities, and tries to match it with the profiles of researchers registered in the system. Fuzzy sets and rough sets endow the system with a kind of intelligence, in order to predict as accurately as possible to what extent a user would classify a document as "interesting".

Figure 1 shows the general architecture of PAS, including the different components and ways of interaction. Below, we discuss some of them in detail.

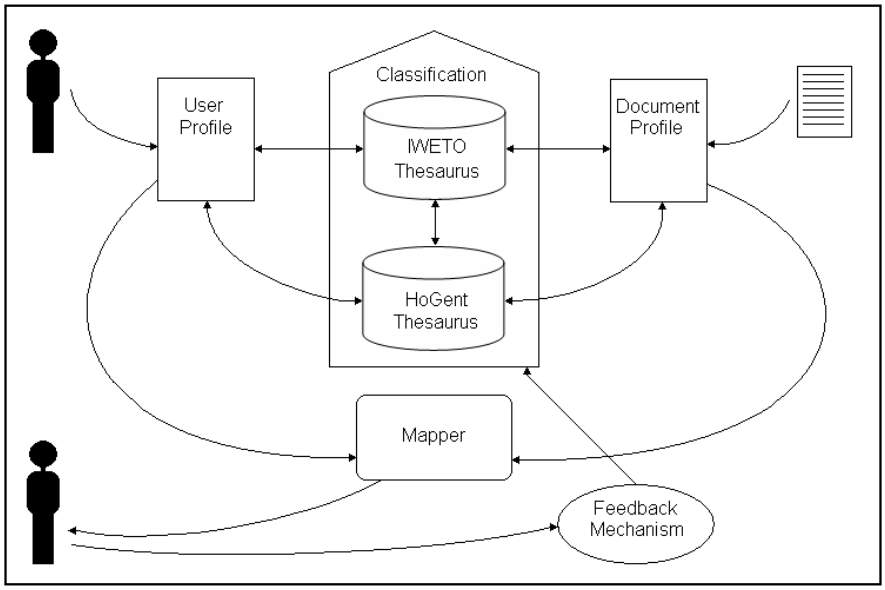

Figure 1: PAS Architecture

3.1 User and document profiles

First of all, the user has to insert his data into the system to create a personal profile that serves as his image in the alert system. The data in the user profile are split into two types: administrative data, and expertise. Administrative data include name, staff number, etc, and are currently not used directly by the system[1]. Expertise refers to data about the research/interest fields of the user, and is used by the system to pinpoint relevant documents. Expertise data are currently entered in the form of two types of keywords: IWETO keywords and free keywords.

[1] In future implementations, we consider using e.g. data about a user's department or research group to enhance the matching process.

IWETO keywords are drawn from the IWETO taxonomy of discipline codes (Iweto 2007), a 3-level tree containing 641 descriptors. The top-level descriptors are Human Sciences, Exact Sciences, Biomedical Sciences, Social Sciences, and Applied Sciences. Each of them is divided into more specific subfields at the second (31) and third (605) level. The latter, however, are still too general to accurately describe a researcher's interest, for example in the third level we can find the field "Programming" but this term is quite generic as it covers a lot of techniques and programming languages. This is why a user is also allowed to enter freely chosen keywords. While these free keywords allow for a higher precision, they also give rise to several problems including misspelling, duplicates, ambiguity, etc. For this reason, currently their selection is limited to a fixed HoGent thesaurus, compiled from employee records.

When selecting keywords, the user can also assign linguistic weights (slightly interested, quite interested, very interested, completely interested) to them. Internally, these linguistic terms are mapped to specific values between 0 and 1, so that every user profile can be seen as a fuzzy set U in the set of keywords.

Information about funding opportunities is stored in the form of "document profiles". Apart from administrative information (contact information, duration, deadline to apply, etc.), these profiles also contain weighted keywords drawn from the IWETO and HoGent thesauri. Currently, the assignment of document keywords is done manually by an administrator, but given the efforts this costs, we plan to automate this task in the future.

3.2 Classification

The classification unites the IWETO and HoGent thesauri into a single graph structure by adding the free keywords as a conceptual fourth layer to the IWETO thesaurus. In this structure, links between terms (keywords) on the third and fourth level are added based on the following principle: if a researcher selects "Artificial Intelligence" as an IWETO keyword to describe his research, and "Java" as a free keyword, a link will be created between these two words in the classification. In this way, a free keyword is likely to have more than one parent in the third layer, and the classification is no longer a tree but a graph.

To reflect that links in the classification can be weaker or stronger, we also weight the graph's links using numbers ranging between 0 (no relation) and 1 (strong relationship). In the current implementation, we assign weights to links based on the co-occurrence of the corresponding keyword pair in user profiles; the more two terms co-occur, the higher the weight associated to their link in the classification.

The classification and its associated weights serve as the basis to construct the fuzzy tolerance relation R that reflects how closely two arbitrary terms are related. This process is illustrated in Figure 2. Formally, let X denote the set of all terms in the classification. A term is of course perfectly related to itself, so $R(t_1,t_1)=1$ for any t_1 in X. If the weight of a path between two terms t_1 and t_2 is defined as the product of its link weights, then $R(t_1,t_2)$ is the maximum weight of all paths between t_1 and t_2, hence $R(t_1,t_2)= max(w1w2, w3w4)$. This also holds when two terms are directly

linked in the classification, so $R(t_1, t_3) = max\ (w1, w3w4w2)$. In Section 3.3, we explain how this classification is used in the matching process.

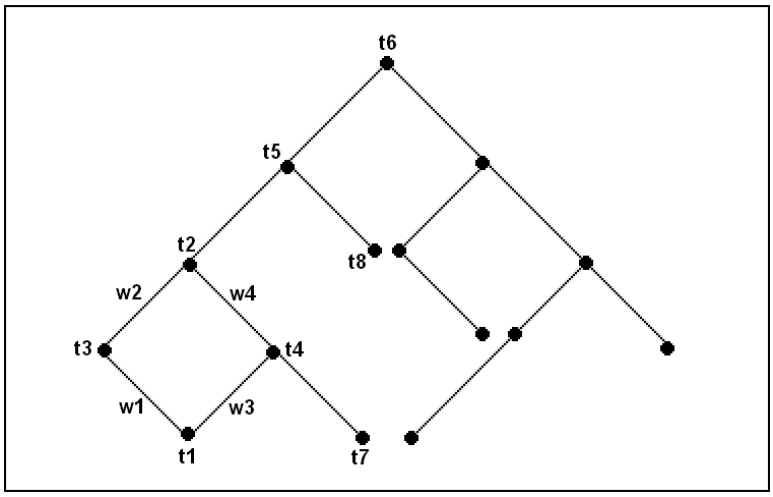

Figure 2: establishing the fuzzy relation R

3.3 The mapper

The mapper is the most important part of the system, since it has to determine whether a document (funding opportunity) is interesting enough to notify a user about it. In this process, we would like the system to be able to identify interesting documents even when their keywords do not exactly match those in the user's profile, but are semantically related to them.

To achieve this added intelligence, we apply fuzzy-rough query expansion. In particular, the system currently uses the approach described in (Srinivasan et al. 2001): to assess how well a document profile D matches a user profile U (both represented as fuzzy sets in the set X of keywords) the algorithm first generates their respective upper approximations[2] $R{\uparrow}D$ and $R{\uparrow}U$, using the fuzzy relation R defined in Section 3.2. The similarity between D and U is then computed, using α-cuts, as

$$Sim_{\alpha}(D, U) = 1 - [\ |B_{u\alpha}| / |\ (R{\uparrow}U)_{\alpha}|\ ]$$

where

$$B_u = R{\uparrow}U\ \ |\!-\!|\ [R{\uparrow}D \cap R{\uparrow}U\ ]$$

[2] (Srinivasan et al. 2001) also considered the lower approximations, but for our purposes this turned out to be too restrictive, as most of the time the lower approximations were empty.

If $|(R\uparrow U)_\alpha|$ is 0, the similarity is defined to be 0. Currently, the system uses a fixed value $\alpha = 0.5$, which turned out to yield the best performance experimentally.

The obtained user-dcoument similarity is then compared to a user-set notification threshold (entered linguistically, just like with the keyword weights, and then mapped onto a numerical value): if the similarity is greater than or equal to the threshold, the system will notify user *U* about document *D*. In this way, the user can tune the sensitivity of the mapper: if he only wants to receive messages that are definitely relevant, he can impose a higher threshold.

3.4 Notification and feedback

Once the user and document profiles are ready, the mapper reads them and tries to find those documents that best fit a given user profile. This process can be configured so that it happens every hour, every day, every week, etc. If the mapper has found any document that could be interesting for the user (based on the criteria described in the previous section), it sends him a notification about it. Optionally, the user can give his feedback about the received notification. This way, users can inform the system whether they are satisfied with the selected messages or not.

One way of taking into account feedback about a notification is to apply changes to the classification, in particular to adjust the term relationships that the notification pertains to. For instance, repeated negative feedback might indicate the need to weaken some links in the classification, while positive feedback could lead to certain relationships being strenghtened. Currently, this process is done manually by an administrator, but we plan to implement an automated feedback mechanism in the foreseeable future.

3.5 An illustrative example

In this section, we give a small example to illustrate how the matching process in PAS works. In Figure 3, two very basic user and document profiles are shown, along with a simplified term classification. The classification is used to construct the fuzzy relation *R* as discussed in Section 3.2.

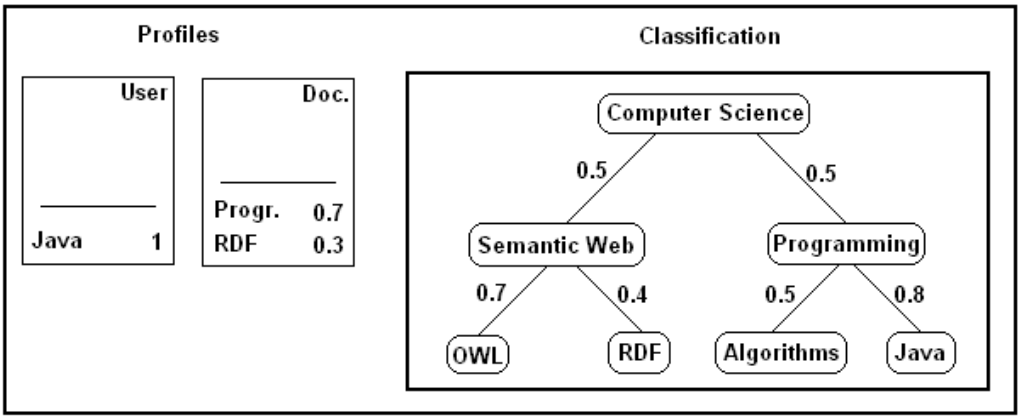

Figure 3: Example

When reading the profiles, the mapper constructs two fuzzy sets:

D (document profile) = {0.7 / *Programming*, 0.3 / *RDF*}
U (user profile) = {1 / *Java*}

Both sets are then compared using the discussed algorithm: first, the membership functions are used to obtain the upper approximations of both sets:

$R{\uparrow}D$ = {0.3/*RDF*, 0.7/*Programming*, 0.5/*Algorithms*, 0.7/*Java,* 0.5 /*CS,* 0.3/*SW,* 0.28/*OWL*}
$R{\uparrow}U$ = {0.08/RDF, 0.8/*Programming*, 0.4/*Algorithms*, 1/*Java,* 0.4/*CS,* 0.2/*SW,* 0.14/*OWL*}

Next, the similarity of D and U is computed. We get, using different values of α:

B_u = {0.3 / Java, 0.1 / Programming}

$Sim_\alpha(D, U) = 1 - 2/6 = 2/3$, if $\alpha = 0.1$
$Sim_\alpha(D, U) = 1 - 1/4 = 3/4$, if $\alpha = 0.3$
$Sim_\alpha(D, U) = 1 - 0/2 = 1$, if $\alpha = 0.5$

In this case, even with a quite high value for the notification threshold, the system will conclude that the user is interested in the project or funding opportunity described by this document. So while the document and user profile actually do not have any keyword in common, the system can exploit the semantical relationships between them to notify the user.

4 Future work

As indicated in some of the previous paragraphs, the current version of PAS is still quite basic, and a lot of work remains to be covered.

For example, a number of tasks are currently done manually, but should be automated if the system is to be deployed on a large scale. This is the case for the feedback mechanism, and also for the creation of the document profiles. The latter is no trivial task, as it requires mapping the documents –which can originate from various sources and do not necessarily adhere to a particular structure or terminology– onto profiles expressed using the classification keywords of the system.

Next, there are several ideas to improve the mapper itself, including using more general definitions for defining fuzzy-rough upper approximation, as well as different similarity measures. We also intend to complement the alert facility with a search engine that allows users to browse the available research information. Note that the same query expansion techniques that are used for personalized notification can also be applied here.

In a similar vein, the mapping algorithms can be applied not only to match users with documents, but also to detect links between researchers in different departments with similar interests, in an effort to boost collaboration.

5 Conclusion

In this paper, we have presented an application of fuzzy rough sets to the field of personal search agents. We have shown how they can be used to provide a flexible user-document mapping mechanism that takes into account semantical links between keywords. More in particular, fuzzy sets are used in the construction of user and document profiles: i.e., they express a researcher's (partial) interest or competence in certain fields, and the relevance of these fields to a document describing a given funding opportunity or project description. Moreover, fuzzy relations are used to represent the gradual relationships between research fields (i.e., the classification). On the other hand, the fuzzy-rough upper approximation is used for expanding user and document profiles with related keywords, and finally a similarity measure is applied to the result in order decide whether to notify the user about the document or not.

6 References

EraCareers (1995): URL: http://ec.europa.eu/eracareers/index_en.cfm

Zadeh, L. A. (1965): Fuzzy Sets. In: *Information and Control* 8 (1965), 338-353.

Pawlak, Z. (1982): Rough Sets. In: *International Journal of Computer and Information Sciences* 11 (1982), 5, 341-356.

De Cock, M.; Cornelis, C.; Kerre E. E. (2007): Fuzzy Rough Sets: the Forgotten Step. In: *IEEE Transactions on Fuzzy Systems* 15 (2007), 1, 121-130.

Magalhães, S. T. (1982): Rough Sets. In: *International Journal of Computer and Information Sciences* 11 (1982), 5, 341-356.

Iweto (2007): URL: http://www.iweto.be

Srinivasan, P.; Ruiz, M. E.; Kraft, D. H.; Chen, J.: Vocabulary mining for information retrieval: rough sets and fuzzy sets. In: *Information Processing and Management*, 37 (2001), 1, 15-38.

Radzikowska, A. M.; Kerre, E. E. (2002): A comparative study of fuzzy rough sets. In: *Fuzzy Sets and Systems* 126 (2002), 137-156.

De Cock, M.; Cornelis, C. (2005): Fuzzy Rough Set Based Web Query Expansion. In: *Proceedings of Rough Sets and Soft Computing in Intelligent Agent and Web Technology, International Workshop at WIIAT2005* (2005), 9-16.

Yao, Y. Y. (1997): Combination of rough and fuzzy sets based on α-level sets. In: *Rough sets and data mining: analysis for imprecise data* (1997), 47-73.

7 Contact Information

Germán Hurtado Martín, Helga Naessens
Dept. of Industrial Sciences, Hogeschool Gent
Schoonmeersstraat 52
9000 Gent
Belgium

e-mail: {German.HurtadoMartin, Helga.Naessens}@HoGent.be
www.cwi.ugent.be/people.php?userid=german

Chris Cornelis
Dept. of Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9)
9000 Gent
Belgium

e-mail: Chris.Cornelis@UGent.be
www.cwi.ugent.be/people.php?userid=chris