A Personalized Information Filtering System for Research Resources based on Multi-Granular Fuzzy Linguistic Modeling

C. Porcel
Dept. of Computing and Numerical Analysis
University of Córdoba, Córdoba, Spain
carlos.porcel@uco.es

E. Herrera-Viedma
Dept. of Computer Science and Artificial Intelligence
University of Granada, Granada, Spain
viedma@decsai.ugr.es

S.Alonso
Dept. of Software Engineering
University of Granada, Granada, Spain
salonso@decsai.ugr.es

A.G. López-Herrera
Dept. of Computer Science
University of Jaén, Jaén, Spain
aglopez@ujaen.es

Abstract

Nowadays, the increasing popularity of Internet has led to an abundant amount of information created and delivered over electronic media. It causes the information access by the users is a complex activity and they need tools, such as information filtering systems to assist them to obtain the required information. Another obstacle is the great variety of representations of information, specially when the users take part in the process, so we need more flexibility in the information processing; the fuzzy linguistic modeling allows to represent and handle flexible information. In this paper, we propose a personalized fuzzy linguistic information filtering system to aid researchers and companies to obtain automatically research resources in their interest areas.

Keywords: information filtering, recommender systems, fuzzy linguistic modeling, technology transfer.

1 Introduction

The Office of Technology Transfer (OTT) is responsible for putting into action and managing the activities which generate knowledge and technical and scientific collaboration, thus enhancing the interrelation between researchers at the University and the entrepreneurial world and their participation in various support programmes designed to carry out research, development and innovation activities. The main mission in this office is to encourage and help, from the University, the generation of knowledge and its spread and transfer in society, with the aim of rapidly meeting society's needs and demands.

The OTT is composed by a team of transfer technology experts that provide to the researchers and companies information about research resources, which could be bulletins, calls, notices, events, congresses, courses and so on. This task requires the selection by the expert of suitable users to deliver the information.

In this paper is proposed SIRE2IN (SIstema de REcomendaciones sobre REcursos de INvestigación - Recommender System about Research Resources) , a fuzzy linguistic information filtering system for research resources. This system is designed to help researchers and companies to find possible collaboration projects, recommending them projects in which they could cooperate. SIRE2IN is designed using both Information Filtering (IF) tools (whose objective is to evaluate and filter the great amount of information available in a specific scope to assist the users in their information access processes) [5, 16] and the multigranular Fuzzy Linguistic Modeling (FLM) to represent and handle flexible information by means of linguistic labels [7, 18].

The paper is structured as follows. Section 2 introduces the IF techniques and the FLM that we use in the system. Section 3 presents

the system SIRE2IN. Finally, we point out our concluding remarks.

2 Preliminaries

2.1 Information Filtering

Information gathering in Internet is a complex activity. Find the appropriate information, required for the users, on the Web is not a simple task. This problem is more acute with the ever increasing use of the Internet. To improve the information access on the Web the users need tools to filter the great amount of information available across the Web. IF is a name used to describe a variety of processes involving the delivery of information to people who need it. It is a research area that offers tools for discriminating between relevant and irrelevant information. IF systems are characterized because they [5, 16] are applicable for unstructured or semi-structured data (e.g. web documents or email messages), are based on user profiles, handle large amounts of data, deal primarily with textual data and their objective is to remove irrelevant data from incoming streams of data items.

Traditionally, these IF systems have fallen into two main categories [5, 16]. *Content-based IF systems* filter and recommend the information by matching the terms used in the representation of user profiles with the terms used in the representation of resources, ignoring data from other users. *Collaborative IF systems* use explicit or implicit preferences from many users to filter and recommend documents to a given user, ignoring the representation of the resources. In this kind of systems, the users' information preferences can be used to define user profiles that are applied as filters to streams of documents; the recommendations to a user are based on another users' recommendations with similar profiles. The construction of accurate profiles is a key task and the system's success will depend on the ability of the learned profiles to represent the user's preferences [15]. Several researchers are exploring hybrid content-based and collaborative IF systems to smooth out the disadvantages of each one of them [4].

Normally, the filtering activity is followed by a relevance feedback phase [15]. Relevance feedback is a cyclic process whereby the user feeds back into the system decisions on the relevance of retrieved documents and the system then uses these evaluations to automatically update the user profile.

Another important aspect that we must have in mind is the method to gather user information, in order to discriminate between relevant and irrelevant information for a user. Information about user preferences can be obtained in two different ways [5], implicit and explicit mode. The *explicit approach*, interacts with the users by acquiring feedback on information that is filtered, that is, the user expresses some specifications of what they desire. The *implicit approach* is implemented by inference from some kind of observation. The observation is applied to user behavior or to detecting a user's environment. The user preferences are updated by detecting changes while observing the user. Moreover, we can combine both approaches in a hybrid approach.

2.2 Fuzzy Linguistic Modeling

There are situations in which the information cannot be assessed precisely in a quantitative form but may be in a qualitative one. For example, when attempting to qualify phenomena related to human perception, we are often led to use words in natural language instead of numerical values. In other cases, precise quantitative information cannot be stated because either it is unavailable or the cost for its computation is too high and an approximate value can be applicable. The use of Fuzzy Sets Theory has given very good results for modeling qualitative information [18] and it has proven to be useful in many problems, e.g., in decision making [8], quality evaluation [13], models of information retrieval [10], clinical decision making [3], political analysis [1], etc. It is a tool based on the concept of *linguistic variable* proposed by Zadeh [18]. Next we analyze the two approaches of FLM that we use in our system.

2.2.1 The 2-Tuple Fuzzy Linguistic Approach

The 2-tuple FLM [7, 9] is a continuous model of representation of information that allows to reduce the loss of information typical of other fuzzy linguistic approaches (classical and ordinal [6, 18]). To define it we have to establish the 2-tuple representation model and the 2-tuple computational model to represent and aggregate the linguistic information, respectively.

Let $S = \{s_0, ..., s_g\}$ be a linguistic term set with odd cardinality, where the mid term represents a indifference value and the rest of the terms are symmetric relate to it. We assume that the semantics of labels is given by means of triangular membership functions and consider all terms distributed on a scale on which a total order is defined, $s_i \leq s_j \Longleftrightarrow i \leq j$.

If a symbolic method used to aggregate linguistic information [6] obtains a value $\beta \in [0, g]$, and $\beta \notin \{0, ..., g\}$, then an approximation function is used to express the result in S. To do this, we represent β by means of 2-tuples (s_i, α_i), where s_i $(i = round(\beta))$ represents the linguistic label, and $\alpha_i = \beta - i$ is a numerical value expressing the value of the translation from the original result β to the closest index label, i, in the linguistic term set. Δ is bijective, that is, $\Delta^{-1}(s_i, \alpha) = \beta \in [0, g]$ [7].

The computational model is defined by presenting the following operators:

1. Negation operator: $Neg((s_i, \alpha)) = \Delta(g - (\Delta^{-1}(s_i, \alpha)))$.
2. Comparison of 2-tuples (s_k, α_1) and (s_l, α_2):
   - If $k < l$ then (s_k, α_1) is smaller than (s_l, α_2).
   - If $k = l$ then
     (a) if $\alpha_1 = \alpha_2$ then (s_k, α_1) and (s_l, α_2) represent the same information,
     (b) if $\alpha_1 < \alpha_2$ then (s_k, α_1) is smaller than (s_l, α_2),
     (c) if $\alpha_1 > \alpha_2$ then (s_k, α_1) is bigger than (s_l, α_2).
3. Aggregation operators: using functions Δ and Δ^{-1} any of the existing aggregation operator can be easily extended for dealing with linguistic 2-tuples, such as arithmetic mean, weighted average or linguistic weighted average.

2.2.2 The Multi-Granular Fuzzy Linguistic Modeling

In any fuzzy linguistic approach, an important parameter to determinate is the "granularity of uncertainty", i.e., the cardinality of the linguistic term set S. According to the uncertainty degree that an expert qualifying a phenomenon has on it, the linguistic term set chosen to provide his knowledge will have more or less terms. When different experts have different uncertainty degrees on the phenomenon, then several linguistic term sets with a different granularity of uncertainty are necessary [8, 12]. The use of different labels sets to assess information is also necessary when an expert has to assess different concepts, as for example it happens in information retrieval problems, to evaluate the importance of the query terms and the relevance of the retrieved documents [11]. In such situations, we need tools for the management of multi-granular linguistic information. In [8] is proposed a multi-granular 2-tuple FLM based on the concept of linguistic hierarchy [2].

A *Linguistic Hierarchy, LH,* is a set of levels *l(t,n(t))*, i.e., $LH = \bigcup_t l(t, n(t))$, where each level t is a linguistic term set with different granularity $n(t)$ from the remaining of levels of the hierarchy [2]. The levels are ordered according to their granularity, i.e., a level $t + 1$ provides a linguistic refinement of the previous level t. We can define a level from its predecessor level as: $l(t, n(t)) \rightarrow l(t + 1, 2 \cdot n(t) - 1)$. In [8] a family of transformation functions between labels from different levels was defined. To define the computational model, we select a level to make uniform the information (for instance, the great granularity level) and then we can use the operators defined in the 2-tuple FLM.

3 SIRE2IN

In this section we present SIRE2IN, a personalized IF system designed using the content-based IF approach and the multi-granular FLM. The system is used to filter the great amount of information that OTT experts manage and spread. SIRE2IN filters the incoming information stream and delivers it to the suitable researchers in accordance with their research areas. For each user the system generates an email with a summary about the resources, its relevance degrees and recommendations about collaboration possibilities.

3.1 System Architecture

It is composed of three main components:

- **Resources management**. It manages the information sources from which the OTT experts receive all the information about research resources. To represent an item, we use the title, abstract, text, date, source, link, kind of resource (project call, events, etc.), target users (researchers, companies or anybody), minimum and maximum financing (for projects) and its scope. To represent the scope we use the *UNESCO terminology* for the science and technology [17]. We use a vector model where for each resource i the system stores a vector VR_i of 248 positions, one position for each discipline. Each position $VR_i[j]$ stores the importance degree for the resource scope i of the UNESCO code represented in j.

- **User profiles management**. The system represents each user through a user profile. To define a user profile we are going to use the identification, contact data, email, research group or company, collaboration preferences (if they want to collaborate with other researchers, a company, anybody or nobody), preferences about the resources (kind, financing, etc.) and topics of interest. The topics of interest are defined by the UNESCO terminology [17] too, i.e. each user chooses a list of UNESCO codes according to his/her information needs or interests. For each user x the system stores a vector VU_x of 248 positions, where each position $VU_x[y]$ stores the importance degree of the UNESCO code represented in the position y for the topics of interest of x.

- **Filtering process**. Based on a matching process the system filters the incoming information to deliver it to the fitting users.

The system uses different labels sets (S_1, S_2,...), chosen from a LH, to represent the different concepts to be assessed. We distinguish three concepts: **importance degrees** of UNESCO codes (S_1), **relevance degrees** of a resource for a user (S_2) and **compatibility degrees** between two users (S_3). Specifically we use labels sets selected of a LH of 3 levels of 3, 5 and 9 labels each one, that is, $S_1 = S^5$, $S_2 = S^9$ and $S_3 = S^3$.

3.2 System Activity

It is based in the following three processes.

3.2.1 Users Insertion Process

This process consists in to incorporate users' data into the system. It presents a form where the users insert their personal information, collaboration preferences and preferences about the resources. Users are invited to define their topics of interest and to choose importance degrees (assessed in S_1) associated with them. Initially a user has associated the topics of interest of his/her research group, but he/she can modify them. The system registers the users and assigns them an identifier (email) and a password. Finally, users receive a confirmation email with the inserted information.

3.2.2 Resources Insertion Process

This process is carried out by the OTT experts that receive or find information about

a resource and they want to spread this information. The experts incorporate the interesting resources into the system and it automatically sends the information to the suitable users along with a relevance degree and collaboration possibilities. When the experts are going to insert a new resource, they insert all the information about it, such as title, abstract, text, date, source, link, kind of resource, users target and minimum and maximum financing. Then they assess, with a linguistic label $s_i^5 \in S_1$, the importance degree of each UNESCO code of level 2 with regard to the resource scope.

3.2.3 Filtering Process

SIRE2IN follows the based-content approach and therefore it filters the information by matching the terms used in the representation of user profiles with the terms used in the representation of resources. As we have said, we use a vector model [14] to represent the resources scope and the user topics of interest. So, to do the matching process we use the cosine angular measure:

$$\sigma(VR, VU) = \frac{\sum_{k=1}^{n}(r_k \times u_k)}{\sqrt{\sum_{k=1}^{n}(r_k)^2} \times \sqrt{\sum_{k=1}^{n}(u_k)^2}}$$

where n is the number of terms (248 in our design), r_k is the value of term k in the resource scope vector and u_k is its value in the user topics of interest vector. With this measure, we obtain a value ranging from 1 for the highest similarity to 0 for the lowest, so we set a threshold value α to filter out the information. Next the system takes into account the user preferences to consider the user or not, and the collaboration preferences. If the user wants to collaborate, the similarity with other users is calculated using the cosine measure too. Finally the system sends to the selected users the resource information, its calculated relevance degree (label of S_2) and the recommendations about collaboration possibilities along with a compatibility degree (label of S_3). To transform labels between different levels of LH, we use the transformation functions defined in the multi-granular FLM.

3.2.4 Feedback Phase

This phase is related to the activity developed by the filtering system once user has taken some of the resources delivered by the system. As we said, user profiles represents the user's long-term information needs or interests and a desirable property for user profiles is that they should be adaptable since user's needs could change continuously. Because of this, the system allows the users to update their profiles to improve the filtering process. To do this, the users access the system and edit their resources preferences, collaboration preferences or their topics of interest, adding or removing UNESCO codes or modifying the importance degree assigned to an existing UNESCO code.

4 Concluding remarks

The exponential increase of Web sites and resources is contributing to that users not being able to find the information they seek in a simple and timely manner. Therefore they are in need of tools to assist them in their information access processes. In this paper we have studied a particular case of information access tools and we have presented SIRE2IN, a IF system based both content-based IF tools and multi-granular FLM. This system helps researchers and companies to obtain automatically information about research resources interesting for both. The system filters the incoming information stream to spread the information to the fitting users and recommends them about collaboration possibilities.

Acknowledgements

This paper has been developed with the financing of SAINFOWEB project, cod. 00602.

References

[1] B. Arfi. Fuzzy decision making in politics: a linguistic fuzzy-set approach (LFSA). *Political Analysis*, 13 (1), 23-56, 2005.

[2] O. Cordón, F. Herrera and I. Zwir. Linguistic modelling by hierarchical systems of linguistic rules. *IEEE Trans-*

actions on Fuzzy Systems, 10 (1), 2-20, 2001.

[3] R. Degani, G. Bortolan. The Problem of Linguistic Approximation in Clinical Decision Making. *Int. J. of Approximate Reasoning*, 2, 143-162, 1988.

[4] N. Good, J.B. Shafer, J.A. Konstan, A. Borchers, B.M. Sarwar, J.L. Herlocker, J. Riedl. Combining collaborative filtering with personal agents for better recommendations. *Proc. of the Sixteenth National Conference on Artificial Intelligence*, 439-446, 1999.

[5] U. Hanani, B. Shapira, P. Shoval. Information Filtering: Overview of Issues, Research and Systems. *User Modeling and User-Adapted Interaction*, 11, 203-259, 2001.

[6] F. Herrera, E. Herrera-Viedma. Aggregation operators for linguistic weighted information. *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems*, 27, 646-656, 1997.

[7] F. Herrera, L. Martínez. A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems*, 8 (6), 746-752, 2000.

[8] F. Herrera, L. Martínez. A model based on linguistic 2-tuples for dealing with multigranularity hierarchical linguistic contexts in multiexpert decision-making. *IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics*, 31(2), 227-234, 2001.

[9] F. Herrera, L. Martínez. The 2-tuple linguistic computational model. Advantages of its linguistic description, accuracy and consistency. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9, 33-48, 2001.

[10] E. Herrera-Viedma. Modeling the retrieval process of an information retrieval system using an ordinal fuzzy linguistic approach. *J. of the American Society for Information Science and Technology*, 52(6), 460-475, 2001.

[11] E. Herrera-Viedma, O. Cordón, M. Luque, A.G. López, A.M. Muñoz. A Model of Fuzzy Linguistic IRS Based on Multi-Granular Linguistic Information. *International Journal of Approximate Reasoning*, 34 (3), 221-239, 2003.

[12] E. Herrera-Viedma, L. Martínez, F. Mata, F. Chiclana. A Consensus Support System Model for Group Decision-making Problems with Multigranular Linguistic Preference Relations. *IEEE Trans. on Fuzzy Systems*, 13 (5), 644-658, 2005.

[13] E. Herrera-Viedma, E. Peis. Evaluating the informative quality of documents in SGML-format using fuzzy linguistic techniques based on computing with words. *Information Processing & Management*, 39(2), 195-213, 2003.

[14] R.R. Korfhage. *Information Storage and Retrieval*. New York: Wiley Computer Publishing, 1997.

[15] L.M. Quiroga, J. Mostafa. An experiment in building profiles in information filtering: the role of context of user relevance feedback. *Information Processing and Management*, 38, 671-694, 2002.

[16] P. Reisnick, H.R. Varian. Recommender Systems. *Special issue of Comm. of the ACM*, 40 (3), 56-59, 1997.

[17] Clasificacin UNESCO. Ministerio de Educación y Ciencia. http://www.mec.es/ciencia/jsp/plantilla.jsp ?area=plan_idi&id=6&contenido=/files/portada.jsp

[18] L.A. Zadeh. The concept of a linguistic variable and its applications to approximate reasoning. Part I. *Information Sciences*, 8, 199-249, 1975. Part II, *Information Sciences*, 8, 301-357, 1975. Part III, *Information Sciences*, 9, 43-80, 1975.