A filtering and recommender system for e-scholars

José M. Morales-del-Castillo* and
Eduardo Peis

Department of Library and Information Science,
University of Granada,
Colegio Máximo de Cartuja s/n 18071, Granada, Spain
E-mail: josemdc@ugr.es
E-mail: epeis@ugr.es
*Corresponding author

Enrique Herrera-Viedma

Department of Computer Science and A.I.,
University of Granada,
c/ Periodista Daniel Saucedo Aranda s/n 18071, Granada, Spain
E-mail: viedma@decsai.ugr.es

Abstract: The way academic community members develop their research activities, access information resources and communicate with each other has dramatically changed with the irruption of the web. Nevertheless, the tools provided by today's web aren't efficient enough to satisfy many of the specific requirements of this new generation of e-scholars. In this paper we present a filtering and recommender system prototype that applies two recommender approaches in order to provide users valuable information about resources and researchers pertaining to domains that completely (or partially) fit their interests. Its main features and elements are enumerated, and an operational example, which illustrates the way the system works, is presented. Additionally, the system has been evaluated and the experimental results reveal a reasonably good performance of the model here proposed.

Keywords: filtering and recommender systems; digital libraries; scholarly; e-scholars.

Reference to this paper should be made as follows: Morales-del-Castillo, J.M., Peis, E. and Herrera-Viedma, E. (2010) 'A filtering and recommender system for e-scholars', *Int. J. Technology Enhanced Learning*, Vol. 2, No. 3, pp.227–240.

Biographical notes: José M. Morales-del-Castillo is an Assistant Professor with the Library and Information Science Department of the University of Granada (Spain).

Eduardo Peis is a Full Professor with the Library and Information Science Department of the University of Granada (Spain).

Enrique Herrera-Viedma is a Senior Lecturer with the Computer Science and Artificial Intelligence of the University of Granada (Spain).

1 Introduction

Traditionally, scholars have to rely on their own knowledge and skills to search, browse and retrieve resources that help them to explore and learn more about a specific knowledge domain using (among others) services such as specialised or academic libraries. Nowadays, due to the irruption of the web as main channel for accessing, retrieving and exchanging scientific information, researchers and academics have changed the way they locate scientific publications (Lawrence and Giles, 1999) and hence they have shifted from mere scholars [which are defined by Kampa (2002) as "individuals involved in advanced learning within a well-defined specialty area who desire in-depth information to support their research and enable the contribution of further ideas, thoughts, theories, and observations"] to scholars that take advantage of the tools provided by the information society (Angehrn et al., 2008), or *e-scholars*.

Nevertheless, the tools that the web actually provides can hardly satisfy these increasingly demanding and specific information requirements of the e-scholars. These needs derive from the scholarly information activities (Palmer and Cragin, 2008) whose aim is the conduct of research and production of scholarship. These activities can be structured in five main categories (Palmer et al., 2009): searching, collecting, reading, writing and collaborating. Particularly in collaborating tasks, where many times work is developed relaying on team based research (Borgman, 2007), it is essential establishing relationships with other colleagues and associates. Nevertheless, this task can be difficult when the research activity implies opening new multidisciplinary lines of investigation, since it is hard to know *what's hot* and *who's in*, in a certain domain out of that of our specialisation (even if both areas are related or close to each other).

Due to this, scholarly libraries in general and digital scholarly libraries in particular (which are traditionally considered as main nodes to access scientific information by the research and scholarly community) must provide their users new value-added services and tools to ease such kind of undertakings.

To achieve this goal, we present in this paper a filtering and recommender system prototype for digital libraries that serves this specific community of users. The system makes available different recommender approaches in order to provide users diverse and valuable information about resources and researchers pertaining to knowledge domains that completely (or partially) fit that which is of his interest. In such a way, users are able to discover implicit social networks where is possible, for example, to find colleagues to form a workgroup (even a multidisciplinary one).

The paper is structured as follows. In Section 2, we briefly discuss the theoretical basis used to develop the prototype (such as Semantic Web technologies, fuzzy linguistic modelling and the recommender approaches supported by the system) and present the main features and elements of the prototype. An operational example of the performance of the prototype is shown in Section 3 and the outcomes of an experiment to evaluate the system are presented in Section 4. Finally some conclusions are pointed out in Section 5.

2 Main features of the prototype

The system here proposed is based on a previous multi-agent model defined by Herrera-Viedma et al. (2007), which has been improved by the addition of new functionalities and services. In a nutshell, our prototype eases users the access to the

information they required by recommending the latest (or more interesting) resources acquired by the digital library, these are represented and characterised by a set of hyperlink lists called *feeds* or *channels* that can be defined using simple mark-up vocabularies, such as atom (Nottingham, 2005) or RSS (*Really Simple Syndication* or *RDF Site Summary*) in any of its multiple versions (Harvard Law, 2004). The structure of these feeds comprises two areas: a first one where the channel is described by a series of basic metadata and another area where different information items (which represent the web resources to be recommended) are defined. The system is developed by applying different fuzzy linguistic modelling approaches [both ordinal (Zadeh, 1975) and 2-tuple based fuzzy linguistic modelling (Herrera and Martinez, 2000) and Semantic Web technologies (Berners-Lee et al., 2001)]. While fuzzy linguistic modelling supplies a set of approximate techniques to deal with qualitative aspects of problems, defining sets of linguistic labels arranged on a total order scale with odd cardinality, Semantic Web technologies allow making web resources semantically accessible to software agents (Hendler, 2001). In such a way, is possible to improve *user-agent* and *agent-agent* interaction and settle a semantic framework where software agents can process and exchange information.

In the next section, we point out some relevant aspects of the theoretical framework used to develop the prototype.

2.1 Semantic Web technologies

The Semantic Web (Berners-Lee et al., 2001) tries to extend the model of the present web using a series of standard languages that enable enriching the description of web resources and make them semantically accessible. To do that, the project is based on two fundamental ideas:

- the semantic tagging of resources, so that information can be *computable* both by humans and computers
- the development of intelligent agents (Hendler, 2001) capable of operating at a semantic level with those resources and infer new knowledge from them (in this way it is possible shifting from keyword search to the retrieval of concepts).

The semantic backbone of the project is the RDF (*Resource Description Framework*) vocabulary (Beckett, 2004) that provides a data model to represent, exchange, link, add and reuse structured metadata of distributed information sources and therefore, make them directly understandable by software agents. RDF structures the information into individual assertions (resource, property and property value triples) and uniquely characterises resources by means of Uniform Resource Identifiers or URI's, allowing agents to make inferences about them using web ontologies (Gruber, 1995, Guarino, 1998) or work with them using simpler semantic structures, like conceptual schemes or thesauri.

As we can see, the Semantic Web basically works with information written in natural language (although structured in a way that can be interpreted by machines). For this reason, it is usually difficult to deal with some problems that require operating with linguistic information that has a certain degree of uncertainty (as, for instance, when quantifying the user's satisfaction in relation to a product or service). A possible solution

could be the use of fuzzy linguistic modelling techniques as a tool for improving the *communication* between system and user.

2.2 2-tuple and ordinal fuzzy linguistic modelling approaches

Fuzzy linguistic modelling (Zadeh, 1975) supplies a set of approximate techniques appropriate to deal with qualitative aspects of problems. The ordinal linguistic approach is defined according to a finite set S of linguistic labels arranged on a total order scale and with odd cardinality (seven or nine tags):

$$\left\{s_i, i \in H = \{0,...,T\}\right\}$$

The central term has a value of 'approximately 0.5' and the rest of the terms are arranged symmetrically around it. The semantics of each linguistic term is given by the ordered structure of the set of terms, considering that each linguistic term of the pair (s_i, s_{T-i}) is equally informative. Each label s_i is assigned a fuzzy value defined in the interval (0,1), that is described by a linear trapezoidal property function represented by the 4-tupla (a_i, b_i, α_i, β_i) (the two first parameters show the interval where the property value is 1.0; the third and fourth parameters show the left and right limits of the distribution). Additionally, we need to define the following properties:

1. – *The set is ordered*: $s_i \geq s_j$ *if* $i \geq j$.
2. – *Negation operator* : $Neg\left(s_i\right) = s_j$, *with* $j = T - i$.
3. – *Maximisation operator* : $MAX\left(s_i, s_j\right) = s_i$ *if* $s_i \geq s_j$.
4. – *Minimisation operator* : $MIN\left(s_i, s_j\right) = s_i$ *if* $s_i \leq s_j$.

Besides, it is necessary to define aggregation operators, such as the *linguistic ordered weighted averaging* (LOWA) operator (Herrera et al., 1996), which are capable to combine linguistic information.

To develop our model, we have also applied another approach to model the linguistic information: the 2-tuple based fuzzy linguistic modelling (Herrera and Martinez, 2000). This approach allows reducing the information loss usually yielded in the ordinal fuzzy linguistic modelling (since information is represented using a continuous model instead of a discrete one) but keeping its straightforward word processing.

In this context, if we obtain a value $\beta \in [0,g]$ and $\beta \notin \{0,\ldots,g\}$ as a result of a symbolic aggregation of linguistic information (Herrera et al., 1996), then we can define an approximation function to express the obtained outcome as a value of the set S.

The fundamental base of this approach is the concept of *symbolic translation* (Herrera and Martinez, 2000). Let β the result of aggregating the indexes of a linguistic term set S. Given $i = round$ (β) and $\alpha = \beta - i$, such that $i \in S[0,g]$ and $\alpha \in [-0.5,0.5)$, then α is what we call *symbolic translation*, i.e., the difference between the information expressed by β and the nearest linguistic label $s_i \in S$.

Therefore, given a linguistic term set $S = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6\}$ and $\beta = 3.3$ as a result of a symbolic aggregation operation, we could represent this value through the linguistic 2-tuple $\Delta(\beta) = (s_3, + 0.3)$.

2.3 Recommender approaches

Traditionally, filtering and recommender systems have been classified into two categories (Popescul et al., 2001): systems that provide recommendations about a specific resource according to the opinions given about that resource by different experts with a profile similar to that of the active user (known as collaborative recommender systems) and systems that generate recommendations according to the similarity of a resource with other resources assessed by the active user (i.e., content-based recommender systems).

In both of them, the likeness between profiles or resources can be measured using different similarity functions such as the Salton's cosine (Salton, 1971) or the Dice coefficient (van Rijsbergen, 1979), just to mention a few. The similarity values are obtained interpreting these functions in a linear way, i.e., the higher the similarity value of a resource or profile is, the more relevant it is to generate a recommendation. This is the traditional recommending approach which fit the needs of the vast majority of e-scholars and *new e-scholars* (i.e., junior researchers and students) to deepen into their knowledge in a specific area and we call it *monodisciplinary* approach.

Nevertheless, as discussed above, it is quite common and almost a need for many researchers to keep the track of new developments and advances in other fields related to their specialisation domain. In this way, it is possible for them to widen their research scope, open new research lines and create multidisciplinary workgroups.

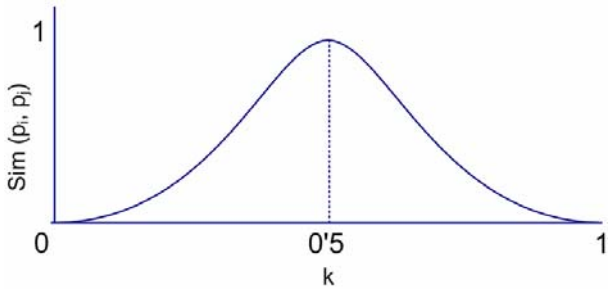
In such circumstance, users require to get recommendations about resources whose topics are related to (but not exactly fit) their preferences but without modifying their profile at all. Therefore, in this case, it makes sense considering as relevant an interval of *mid-range* similarity values instead of those close to one (i.e., both extremely similar and dissimilar similarity values are discarded).

So it would be necessary defining some kind of centre function (Yager, 2007) that enable constraint the range of similarity values we are going to consider as relevant. In our model, the interpretation of similarity is defined by a Gaussian function μ as the following:

$$\mu\left(Sim\left(p_i, r_j\right)\right) = e\left(Sim\left(p_i, r_j\right) - k\right)^2$$

where *Sim* (p_i, p_j) is the similarity measure among the resources p_i and p_j and k represents the centre value around which similarity is relevant to generate a recommendation (in this case $k = 0.5$). This is what we call *multidisciplinary* approach.

Figure 1 Gaussian centre function (see online version for colours)



2.4 Architecture and modules

To carry out the filtering and recommendation process we have defined three software agents (interface, task and information agents) that are distributed in a five level hierarchical architecture:

Level 1 *User level*: in this level users interact with the system by defining their preferences, providing feedback to the system, etc.

Level 2 *Interface level*: this is the level defined to allow interface agent developing its activity as a mediator between users and the task agent. It is also capable to carry out simple filtering operations on behalf of the user.

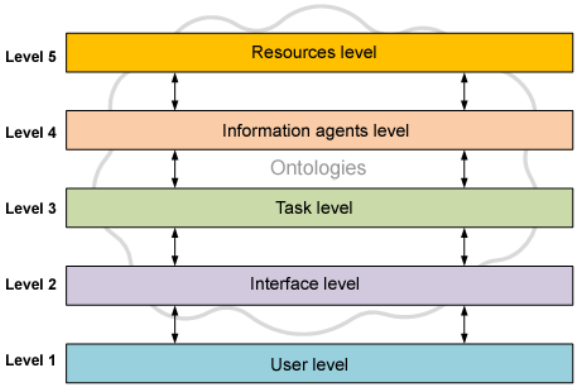
Level 3 *Task level*: in this level is where the task agent (normally one per interface agent) carries out the main load of operations performed in the system such as the generation of information alerts or the management of profiles and RSS feeds.

Level 4 *Information agents level*: here is where several information agents can access system's repositories, thus playing the role of mediators between information sources and the task agent.

Level 5 *Resources level*: in this level are included all information sources the system can access such as a full-text documents repository and a set of resources described using RDF-based vocabularies (RSS feeds containing the items featured by the digital library, a user profile repository and a thesaurus that describes the specialisation domain of the library).

The underlying semantics of the different elements that make up the system (i.e., their characteristics and the semantic relations defined among them) is defined through several interoperable web ontologies described using the OWL vocabulary (McGuinness and van Harmelen, 2004). Since the communication processes carried out among agents in this model involves natural language information and fuzzy linguistic tags, we have chosen to use the adaptation of the FIPA agent communication language (Foundation for Intelligent Agents, 2009) proposed by Willmott et al. (2005), which is based on XML syntax and RDF/OWL as content language.

Figure 2 Levels of the filtering and recommender system (see online version for colours)



In the prototype there are also defined five main activity modules:

- *Profiles generation module*: in this module, users are able to characterise their profiles by defining personal data and a set of concepts that define their long-term information needs. These concepts are lexically matched with the terms of the thesaurus using as similarity measure the edit tree algorithm (Levenshtein, 1966) which compares character strings and returns the same term introduced when there is an exact match, or a term lexically similar to the given term if there is no exact match. In this later case, the system has a mechanism that allows the user to browse the thesaurus' categories and select concepts by himself. Then, each selected concept must also be weighted using a fuzzy linguistic label that represents the degree of interest of the user about that specific topic. In addition, the system adds further filtering levels improving the profile representation through the definition of the stereotypes junior and senior researchers, thus differencing between knowledge levels or skills. In such a way, it is possible to provide different recommendations to users according to the opinion of users with similar interests and knowledge level.
- *RSS generation module*: in this module system administrators or site managers can create and update the RSS feeds of the system in a semi-automatic way through an interface where they can input the different elements needed to describe both the RSS channel and its items. The description of the channel is static (i.e., not susceptible to changes) and includes a title, a brief summary of the content and frequency with which items are updated. Description of the items is continually renewed, deleting out-of-date items and adding new ones according to the updating frequency defined in the channel description. To do so, the task agent periodically checks the document repository seeking for documents that have not been described yet. Once these documents have been located, information agents are responsible for extracting the data needed to generate their description from a web information source (such as, for instance, a database or a public access repository). Then, the task agent proceeds to generate the description of the items by defining a title, an author, a content summary and a link to the primary resource. If the data provided by information agents is wrong or incomplete, system managers are responsible for correcting or completing them. Nevertheless, there must always be a careful human supervision (carried out by system managers) of the assignment of topics terms that describe the content of any resource. To ease this task, we use a tool that helps in the process of assigning topics to the items. It works in an analogous way to the preference selection process in the *profiles generation module*: the administrator suggests a series of terms which are matched with the terms of the thesaurus using the edit tree algorithm and the matched terms will be assigned as topic terms. Here, the system suggests a series of lexically similar terms that site managers can use or not, depending on their own criterion.
- *Information push module*: this module is responsible for generating and managing the information alerts to be provided to users (so it can be considered as the service core). The similarity between user profiles and resources is measured according to the hierarchical lineal operator defined by Oldakowsky and Byzer (2005) which takes into account the position of the concepts to be matched in a taxonomic tree. Once defined this similarity value, the relevance of resources or profiles is calculated according to do the concept of *semantic overlap*. This concept tries to ease the

problem of measuring similarity using taxonomic operators, since all the concepts in a taxonomy are related in a certain degree and therefore, the similarity between two of them would never reach 0 (i.e., we could find relevance values higher than one that can hardly be normalised). The underlying idea in this concept is determining areas of maximum semantic intersection between the concepts in the thesaurus of the system. To obtain the relevance of profiles to other profiles we define the following function:

$$Sim\left(P_i, P_j\right) = \frac{\sum_{k-1}^{MIN(N,M)} H_k\left(Sim\left(\alpha_i, \delta_j\right)\right)\left(\frac{\omega_i + \omega_i}{2}\right)}{MAX\left(N, M\right)}$$

where $H_k(\text{Sim}\ (\alpha_i, \delta_j))$ is a function that extracts the *k* maximum similarities defined between the preferences of $P_i = \{\alpha_{1, \ldots,} \alpha_N\}$ and $P_j = \{\delta_{1, \ldots,} \delta_M\}$ and ω_i, ω_j are the corresponding associated weights to α_i and δ_j. When matching profiles $P_i = \{\alpha_{1, \ldots,} \alpha_N\}$ and items $R_j = \{\beta_1, \ldots, \beta_M\}$, since subjects are not weighted, we will take into account only the weights associated to preferences so the function in this case is slightly different:

$$Sim\left(P_i, R_j\right) = \frac{\sum_{k-1}^{MIN(N,M)} H_k\left(Sim\left(\alpha_i, \beta_j\right)\right)\omega_i}{MAX\left(N, M\right)}$$

- *Feedback or user profiles updating module*: in this module the updating of user profiles is carried out according to users' assessments about the set of resources recommended by the system. This updating process consists in recalculating the weight associated to each preference and adding new entries to the recommendations log stored in every profile. We have defined a matching function that rewards those preference values that are present in resources positively assessed by users and penalised them, on the contrary, when this assessment is negative. Let $e_j \in S'$ be the degree of satisfaction provided by the user, and $\omega^j_{il} \in S$ the weight of property *i* (in this case i=«Preference») with value l. Then, we define the following updating function g: S'xS→S:

$$g\left(e_j, \omega^j_{li}\right) = \begin{cases} s_{Min(a+\beta,T)} & if\ s_a \le s_b \\ s_{Max(0,a-\beta)} & if\ s_a > s_b \end{cases}$$

$$s_a, s_b \in S \mid a, b \in H = \{0, \ldots T\}$$

where

1 $s_a = \omega^j_{li}$

2 $s_b = e_j$

3 *a* and *b* are the indexes of the linguistic labels which value ranges from 0 to T (being T the number of labels of the set S minus one)

4 β is a bonus value which rewards or penalise the weights of the preferences. It is defined as $\beta = round(2|b\text{-}a|/T)$ where *round* is the typical round function.

- *Collaborative recommendation module*: the aim of this module is generating recommendations about a specific resource in base to the assessments provided by

different experts with a profile similar to that of the active user. The different recommendations (expressed through linguistic labels) are aggregated using the LOWA operator. It also allows users to explicitly know the identity and institutional affiliation data of these experts in order to contact them for any scholarly purpose. This feature of the system implies a total commitment between the digital library and its users since their altruistic collaboration can only be achieved by granting that their data will exclusively be used for contacting other researchers subscribed to the library. Therefore, becomes a critical issue defining privacy policies to protect those individuals that prefer to be *invisible* for the rest of users. Nevertheless, we have to point out that this functionality is still in development and has not been implemented yet.

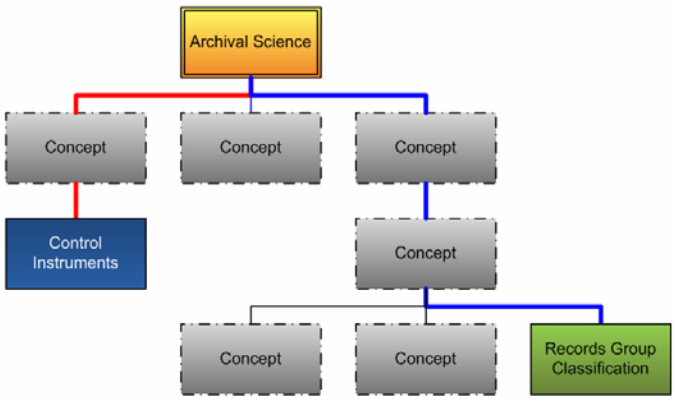
3 Operational example

To clarify the performance of the system, we have developed this operational example. Let's start defining a set of premises:

- a generic user that wants to obtain *monodisciplinary* recommendations from the system, with a profile P where preferences α_1, α_2 (N = 2) and their associated weights ω_1, ω_2 are defined
- an item R of the RSS feed represented by the subjects β_1, β_2, β_3 (M = 3).

First of all the system proceeds to calculate the similarity between the resources in the RSS feed and the profile of the active user applying the taxonomic linear operator defined in (Oldakowsky and Byzer, 2005). Let α_1 be the concept c*ontrol instruments* with a depth of two in the thesaurus of the system and β_2 the concept *record group classification* with a depth of three (being six as the maximum depth of the thesaurus). The closest common parent (*ccp*) of both concepts is *archival science*, which depth is zero by default. As a result, the distance between α_1 and β_2 is $d(\alpha_1, \beta_2) = 0.83$.

Figure 3 Sample concepts in the thesaurus (see online version for colours)



The rest of distances and corresponding similarities are respectively shown on Tables 1 and 2:

Table 1 Distances between preferences and subject concepts

Preferences/subjects	β_1	β_2	β_3
α_1	0.21	0.83	0.12
α_2	0.16	0.07	0.35

Table 2 Similarities between preferences and subject concepts

Preferences/subjects	β_1	β_2	β_3
α_1	0.79	0.17	0.88
α_2	0.84	0.93	0.65

In the next step, the relevance of the item *R* to the profile *P* is calculated. Let the importance value for the preference α_1 be the linguistic label *very high* (i.e., $\omega_1 = 0.83$) and for α_2 the label *medium* (i.e., $\omega_2 = 0.5$). Besides, if the number of preferences and subjects is respectively N = 2 and M = 3, then the three maximum similarities are chosen to calculate the relevance value (in this case, let's suppose Sim (α_1, β_3) = 0.88, Sim (α_2, β_1) = 0.84, and Sim (α_2, β_2) = 0.93). The resulting relevance value is *Rel* (*P*, *R*) = 0.54 so, as the relevance threshold has been fixed in $k = 0.50$, the resource *R* is selected to be retrieved.

Applying the 2-tuple based fuzzy linguistic modelling approach, relevance is displayed as linguistic label extracted from the linguistic variable *relevance level* together with a numeric value: *medium* + 0.04 (i.e., *medium* is the closest label to the relevance value 0.54 and the corresponding numeric value of this label has been exceeded by 0.04).

The following step consists in searching profiles (similar to the profile of the active user) with recommendations about the resource *R* in order to generate a collaborative recommendation. Supposed two users that have respectively assessed the resource *R* with the linguistic labels *high* and *medium* (which have been extracted from the linguistic variable *level of satisfaction*), when applying the LOWA operator (Herrera et al., 1996) the resulting aggregated label is the following: k = MIN{6,3 + round (0.4*(4–3))} =3. Then l_k = *medium*.

As the non-weighted average similarity of the preference α_1 (with a value of 0.80) is lower than that of α_2 (with a value of 0.88), this last preference value will be the chosen to be updated. Let's see an example of the updating process.

Supposed the user assesses the resource *R* (which has satisfied his information needs) defining a satisfaction level with the linguistic label e_j = *very high* (where $e_j \in S'$ = *{null, very low, low, medium, high, very high, total}*). In this case, the associated weight to α_2 is $\omega^j_{(Preference,\ \alpha 2)}$ = *medium* (where $\omega^j_{li} \in S$ = *{null, very low, low, medium, high, very high, total}*). Considering that $s_a \leq s_b$, whose index values are $a = 3$ and $b = 5$ and $T = 6$, we have that $\beta = 1$, so the new associated weight for α_2 is increased in a factor of one $(\omega^j_{(Preference,\ \alpha 2)})' = g$ (*Very high, Medium*) = *high*.

If the user decides to get multidisciplinary recommendations, the process is carried out in a slightly different manner. Let *R* and *R*' be the set of retrieved resources with relevance values *Rel(P, R)* = 0.57 and *Rel(P, R')* = 0.83 respectively the system recalculates both relevance values according to the centring function: μ(Rel (P, R)) = 1.005; μ(Rel (P, R')) = 1.110.

Then, the system rearranges the retrieved items and considers as more relevant the values which are closer to one (in this case, *R* is more relevant than *R*').

4 Prototype evaluation

We have set up an experiment to evaluate the content-based module of the prototype in terms of precision (Cao and Li, 2007) and recall (Cleverdon et al., 1966) [since the collaborative recommendation module is not fully implemented yet and suffers from *cold start problem* (Schein et al., 2002)]. These two measures (together with the F1 measure (Sarwar et al., 2000) are usually used in filtering and recommender systems to assess the quality of the set of retrieved resources.

To carry out the evaluation and according to users' information needs, the set of items recommended by the system have been classified into four basic categories: relevant suggested items (Nrs), relevant non-suggested items (Nrn), irrelevant suggested items (Nis) and irrelevant non-suggested items (Nin). We have also defined other categories to represent the sum of selected items (Ns), non-selected items (Nn), relevant items (Nr), irrelevant items (Ni) and the whole set of items (N).

Based on to these categories we have defined in our experiment precision, recall and F1 as follows:

- *Precision*: ratio of selected relevant items to selected items, i.e., the probability of a selected item to be relevant:

  $P = Nrs/Ns$

- *Recall*: ratio of selected relevant items to relevant items, i.e., the probability of a relevant item to be selected:

  $R = Nrs/Nr$

- *F1*: combination metric that equals both the weights of precision and recall:

  $F1 = (2*P*R*)/(P+R)$.

The goal of the experiment is to test the performance of our prototype in the generation of accurate and relevant content-based recommendations for the users of the system, exclusively considering the mono-disciplinary search. To do so, we have asked a random sample of 12 researchers in the field of library and information science that develop their activity at the University of Granada to evaluate the results provided by the prototype.

One of the premises of the experiment is that at least one of the topics defined for a relevant resource and one of the experts' preferences must be semantically constraint to the same sub-domain of the thesaurus. In such a way, we can leverage a better terminological control on subjects and preferences and extrapolate the output data to the whole thesaurus. In this case, the sub-domain selected is *archival science*, which is composed of 96 different concepts. We also require two more elements:

- an RSS feed that contains 30 items extracted from the E-LIS open access repository (ELIS, 2009), from which only ten of them are semantically relevant (i.e., with at least one subject pertaining to the selected sub-domain)
- a set of user profiles with at least one preference pertaining to the targeted sub-area.

The prototype is set to recommend up to ten resources and then users are asked to assess the results by explicitly stating which of the recommended items they consider are

relevant. With these starting premises the experiment was carried out and the results are shown in Table 3:

Table 3 Experimental data

	User 1	*User 2*	*User 3*	*User 4*	*User 5*	*User 6*	*User 7*	*User 8*	*User 9*	*User 10*	*User 11*	*User 12*
Nrs	6	5	3	6	4	5	5	4	6	3	7	6
Nrn	2	3	2	1	2	3	2	2	2	2	1	2
Nis	4	5	7	4	6	5	5	6	2	7	3	4
Nr	8	8	5	7	6	8	7	6	8	5	8	8
Ns	10	10	10	10	10	10	10	10	10	10	10	10

Precision, recall and F1 for each user are shown in Table 3 (in percentage) and represented in the graph in Figure 4. The average outcomes reveal a quite good performance of the prototype.

Table 4 Detailed experimental outcomes

%	*User 1*	*User 2*	*User 3*	*User 4*	*User 5*	*User 6*	*User 7*	*User 8*	*User 9*	*User 10*	*User 11*	*User 12*	*Ave.*
P	60.00	50.00	30.00	60.00	40.00	50.00	50.00	40.00	60.00	30.00	70.00	60.00	50.00
R	75.00	62.50	60.00	85.71	66.67	62.50	71.43	66.67	75.00	60.00	87.50	75.00	70.66
F1	66.67	55.56	40.00	70.59	50.00	55.56	58.82	50.00	66.67	40.00	77.78	66.67	58.19

Figure 4 Precision, recall and F1 (see online version for colours)

5 Conclusions

In this paper we have presented a multi-agent filtering and recommender system prototype for digital libraries designed to be used by the e-scholars community that provides an integrated solution to minimise the problem of access relevant information in vast document repositories.

The prototype combines Semantic Web technologies and several fuzzy linguistic modelling techniques to define a richer description of information, thus improving communication processes and user-system interaction.

The system is able to generate both *monodisciplinary* recommendations (to deepen into users' specialisation area) and *multidisciplinary* recommendations, which allow users eliciting resources whose topics are tangentially related to their preferences. Furthermore, the model provides additional levels of filtering through the definition of stereotypic profiles according to users' knowledge or skills.

The prototype makes possible for researchers to uncover implicit social networks, which relate them with other researchers from different domains, thus easing the task of forming multidisciplinary working groups. Nevertheless, this implies that the system should apply privacy policies to protect those individuals that prefer to be *invisible* for the rest of users.

The system has been evaluated and experimental results show that the system is reasonably effective in terms of precision and recall, although further detailed evaluations may be necessary.

Acknowledgements

This work has been supported by FEDER funds in the National Spanish Projects TIN2007-61079, PET2007_0460 and FOMENTO-90/07.

References

Angehrn, A.A., Maxwell, K., Luccini, A.M. and Rajola, F. (2008) 'Designing collaborative learning and innovation systems for education professionals', in M.D. Lytras, J.M. Carroll, E. Damiani and R.D. Tennyson (Eds.): *Emerging Technologies and Information Systems for the Knowledge Society. Lecture Notes in Computer Science*, Vol. 5288, pp.167–176, Springer.

Beckett, D. (Ed.) (2004) 'RDF/XML syntax specification', available at: http://www.w3.org/TR/rdf-syntax-grammar/.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The Semantic Web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities', *The Scientific American*, May, available at: http://www.sciam.com/article.cfm?id=the-semantic-web.

Borgman, C.L. (2007) *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*', MIT Press.

Cao, Y. and Li, Y. (2007) 'An intelligent fuzzy-based recommendation system for consumer electronic products', *Expert Systems with Applications*, Vol. 33, No. 1, pp.230–240.

Cleverdon, C.W., Mills, J. and Keen, E.M. (1966) *Factors Determining the Performance of Indexing Systems*, Vol. 2, Test results, ASLIB Cranfield Project.

ELIS (2009) http://eprints.rclis.org/.

Foundation for Intelligent Agents (FIPA) (2009) http://www.fipa.org.

Gruber, T.R. (1995) 'Toward principles for the design of ontologies used for knowledge sharing', *International Journal of Human-Computer Studies*, Vol. 43, Nos. 5–6, pp.907–928.

Guarino, N. (1998) 'Formal ontology and information systems', in N. Guarino (Ed.): *Formal Ontology in Information Systems*, pp.3–17, IOS Press.

Hendler, J. (2001) 'Agents and the Semantic Web', *IEEE Intelligent Systems*, March–April, pp.30–37.

Herrera, F. and Martinez, L. (2000) 'A 2-tuple fuzzy linguistic representation model for computing with words', *IEEE Transactions on Fuzzy Systems*, Vol. 8, No. 6, pp.746–752.

Herrera, F., Herrera-Viedma, E. and Verdegay, J.L. (1996) 'Direct approach processes in group decision making using linguistic OWA operators', *Fuzzy Sets and Systems*, Vol. 79, No. 2, pp.175–190.

Herrera-Viedma, E., Peis, E., Morales-del-Castillo, J.M. and Anaya, K. (2007) 'Improvement of web-based service information systems using fuzzy linguistic techniques and Semantic Web technologies', in J. Liu, D. Ruan and G. Zhang, (Eds.): *E-Service Intelligence: Methodologies, Technologies and Applications*, pp.647–666, Springer Verlag.

Kampa, S. (2002) 'Who are the experts? E-scholars in the Semantic Web', Thesis, Department of Electronics and Computer Science, University of Southampton.

Lawrence, S. and Giles, C.L. (1999) 'Searching the web: general and scientific information access', *IEEE Communications*, Vol. 37, No. 1, pp.116–122.

Levenshtein, V.I. (1996) 'Binary codes capable of correcting deletions, insertions, and reversals', Soviet Physics Doklady, Vol. 10, No. 8, pp.707–10.

McGuinness, D.L. and van Harmelen, F. (Eds.) (2004) 'OWL web ontology language overview', available at: http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Nottingham, M. (2005) 'The atom syndication format', available at: http://atomenabled.org/developers/syndication/atom-format-spec.php.

Oldakowsky, R. and Byzer, C. (2005) 'SemMF: a framework for calculating semantic similarity of objects represented as RDF graphs', available at: http://www.corporate-semantic-web.de/pub/SemMF_ISWC2005.pdf.

Palmer, C.L. and Cragin, M.H. (2008) 'Scholarly information work and disciplinary practices', *Annual Review of Information Science and Technology*, Vol. 42, pp.165–211.

Palmer, C.L., Teffeau, L.C., Pirmann, C.M. (2009) 'Scholarly information practices in the online environment: themes from the literature and implications for library service development', Report commissioned by OCLC Research, available at: http://www.oclc.org/programs/publications/reports/2009-02.pdf.

Popescul, A., Ungar, L.H., Pennock, D.M. and Lawrence, S. (2001) 'Probabilistic models for unified-collaborative and content-based recommendation in sparse-data environments', in J.S. Breese and D. Koller (Eds.): *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, Morgan Kaufmann pp.437–44.

RSS 2.0 at Harvard Law (2004) http://cyber.law.harvard.edu/rss/rssVersionHistory.html.

Salton, G. (1971) 'The Smart retrieval system – experiments', in G. Salton (Ed.): *Automatic Document Processing*, Prentice–Hall, Englewood Cliffs.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2000) 'Analysis of recommendation algorithms for e-commerce', in A. Jhingran, J.M. Mason and D. Tygar (Eds.): *Proceedings of ACM E-Commerce 2000 Conference*, pp.158–167, ACM.

Schein, A.I., Popescul, A. and Ungar, L.H. (2002) 'Methods and metrics for cold-start recommendations', in K. Jarvelin, M. Beaulieu, R. Baeza-Yates and S.H. Myaeng (Eds.): *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Developmenting Information Retrieval (SIGIR 2002)*, pp.253–260, ACM.

van Rijsbergen, C.J. (1979) *Information Retrieval*, Butterworths.

Willmott, S., Peña, F.O.F., Merida-Campos, C., Constantinescu, I., Dale, J. and Cabanillas, D. (2005) 'Adapting agent communication languages for Semantic Web service inter-communication', in A. Skowron et al. (Eds.): *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence 2005*, pp.405–408, IEEE Society.

Yager R.R. (2007) 'Centered OWA operators', *Soft Computing*, Vol. 11, No. 7, pp.632–639.

Zadeh, L.A. (1975) 'The concept of a linguistic variable and its applications to approximate reasoning', *Information Sciences*, Vol. 8, No. 1, pp.199–249; Vol. 8, No. 2, pp.301–357; Vol. 9, No. 3, pp.43–80.