2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology

A web-based service for the elicitation of resources in the biomedical domain

J.M. Morales-del-Castillo
Dpt. Library and Information Science,
University of Granada,
Granada, Spain,
e-mail: josemdc@ugr.es

Carlos Porcel
Dpt. Computer Science,
University of Jaén
Jaén, Spain
e-mail:cporcel@ujaen.es

E. Herrera-Viedma
Dpt. Computer Science and A.I.,
University of Granada,
Granada, Spain
e-mail:viedma@decsai.ugr.es

Eduardo Peis
Dpt. Library and Information Science,
University of Granada,
Granada, Spain
e-mail: epeis@ugr.es

***Abstract*—In certain domains with a dynamic research activity, such as that of Biomedical Sciences, it is necessary the development of new services capable of satisfying their specific information needs. In this paper we present a filtering and recommender system that applies Semantic Web technologies and Fuzzy Linguistic Modeling techniques in order to provide users valuable information about resources that fit their interests. The main features and elements of the system are enumerated in this paper, and an operational example (which illustrates the overall system performance) is presented. Furthermore, the outcomes of a simple system evaluation are shown.**

Keywords—Filtering and Recommender Systems, Biomedical Sciences, Semantic Web technologies, Fuzzy Linguistic Modeling.

I. INTRODUCTION

Nowadays, one of the main challenges that information systems have to face is the efficient management of resources in order to satisfy the increasingly more complex and specific requirements of their users. In dynamic and very productive domains, such as Biomedical Sciences (where the vast majority of the knowledge that is generated is published in the form of scientific papers [17]), information overload is big handicap to accessing relevant resources since it is a hard task (and virtually impossible) for a biomedical researcher trying to keep up with the latest researching trends and breakthroughs on his/her specialty (even more when the level of granularity of their information needs is so high).

Current web services have shown their inability to provide an accurate and efficient response to these requirements, since information in the Web is basically represented using natural language, and machines aren't capable to interpret and contextualize it. Therefore, it is becoming necessary to develop systems for searching and mining the Web that allow improving the access to information in an efficient way. At this moment, some of the more recurrent technologies to face this problem deal with the development of intelligent software agents [5], the application of techniques of information filtering [22], and the development of knowledge-based applications using Semantic Web technologies (such as the Biogateway Portal [2] or the *National Cancer Institute Thesaurus* [14]).

Nevertheless the main problem of using agents is to find a flexible and agile communication protocol for exchanging information among agents, and between users and agents because of the great variety of forms the information is represented in the Web. A possible option that permits to reduce these agent-agent and user-agent communication problems is to apply fuzzy linguistic techniques that allow operating with the information by means of the use of linguistic labels [23]. The application of this flexible system of representation enables us to handle information with several degrees of truth, solving the problem of quantifying qualitative concepts.

Our proposal is the development of multi-agent filtering and recommender system that jointly applies Semantic Web technologies and Fuzzy Linguistic Modeling techniques to provide biomedical researchers a better access to resources of their interest.

The paper is structured as follows. In section 2 we briefly discuss the theoretical basis used to develop the system (such as Semantic Web technologies and Fuzzy Linguistic Modeling) and present the main features and elements of the system. The structure and modules of the system are 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.

II. THEORETICAL BASIS

The system here proposed is based on a previous multi-agent model defined by Herrera-Viedma et al. [11], which has been improved by the addition of new functionalities and services. In a nutshell, our system eases users the access to specialized information they required by recommending the latest (or more interesting) resources published in a specific domain (in this case, biomedicine). These resources 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 [15] or RSS (*Really Simple Syndication* or *RDF Site Summary* as well) in any of its multiple versions [19]. 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 containing different information items that represent the web resources to be recommended.

The system is developed by the application of Semantic Web technologies [1] [8] to improve *user-agent* and *agent-agent* interaction, and to settle a semantic framework where software agents can process and exchange information using Web ontologies [6][7] (or simpler semantic structures like conceptual schemes or thesauri), and fuzzy linguistic modeling techniques [23], which allow dealing 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). Among these techniques and tools we can find a diversity of aggregation operators, such as the *Linguistic Ordered Weighted Averaging* (LOWA) operator [9], which are capable to combine linguistic information. In

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.319

this context, the 2-tuple based fuzzy linguistic modelling [10], where information is represented using a continuous model instead of a discrete one. It is based on the concept of "*symbolic translation*", defined as the difference between the information expressed by an aggregated linguistic value β and the nearest linguistic label s_i in the set of possible linguistic values to describe a specific dimension. Therefore, any value expressed according to this approach will be defined by a linguistic tag and a number that represents a positive or negative "*symbolic translation*" (for instance, "*High* + 0.3").

III. Structure and Modules of the System

To carry out the filtering and recommendation process we have defined 3 software agents (interface, task and information agents) that are distributed in a 5 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 the information sources the system can access: a document repository (in this case we have opted for using the public database PubMed [18]), a set of RSS feeds containing the items to be recommended, a user profile repository and a test thesaurus in SKOS [12] format, that has been developed taking as a model the *National Cancer Institute Thesaurus* [14].

The underlying semantics of the different elements that make up the system (i.e. their characteristics and the semantic relations defined among them) are defined through several interoperable web ontologies described using the OWL vocabulary [13].

In the system there are also defined 3 main activity modules:

- *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 [16] which takes into account the position of the concepts to be matched in a taxonomic tree. Once the similarity between preferences and topic terms is defined, the relevance of resources or profiles is calculated according to 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 1 that can hardly be normalized). The underlying idea in this concept is determining areas of maximum semantic intersection between the concepts in the taxonomy. To obtain the relevance of profiles according to other profiles we define the following function:

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

where $H_k(Sim\ (\alpha_i,\ \delta_j))$ is a function that extracts the *k* maximum similarities defined between the preferences of $P_i = \{\alpha_{1,\ ...,}\ \alpha_N\}$ and $P_j = \{\delta_{1,\ ...,}\ \delta_M\}$, and ω_i, ω_j are the corresponding associated weights to α_i and δ_j. When matching profiles $P_i = \{\alpha_{1,\ ...,}\ \alpha_N\}$ and items $R_j = \{\beta_{1,\ ...,}\ \beta_M\}$, 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:

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

- *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 penalized 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:

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

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

where, (i) $s_{a=}\ \omega^{j}_{li}$; (ii) $s_{b=}\ e_j$; (iii) *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), and (iv) βis a bonus value which rewards or penalize the weights of the

preferences. It is defined as $\beta=round(2|b-a|/T)$ 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 [9] and displayed according to the 2-tuple based fuzzy modelling approach [10]. The system also allows users to explicitly know the identity and institutional affiliation data of these experts in order to contact them for any research purposes. This feature of the system implies a total commitment between the service 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 service. 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.

IV. Evaluation of the System

We have set up an experiment to evaluate the content-based module of the system in terms of precision [3] and recall [4] (since the collaborative recommendation module is not fully implemented yet and suffers from *cold start problem* [21]). These two measures (together with the F1 measure [20] 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 system 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 ten researchers in the field of Biomedicine to evaluate the results provided by the system.

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 "*Angiogenesis Inhibitor*", which is composed of around 100 different concepts. We also require two more elements:

- an RSS feed that contains 30 items extracted from the PubMed repository [18], from which only 10 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 system is set to recommend up to 10 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 1:

TABLE 1

Precision, recall and F1 for each user are shown in table 2 (in percentage) and represented in the graph in figure 1. The average outcomes reveal a quite good performance of the system (nearly close to the 50% in terms of precision).

TABLE 2

FIG 1

V. Conclusions

In this paper we have presented a multi-agent filtering and recommender system (designed to be used by biomedical researchers) which provides an integrated solution to minimize the problem of access relevant information in vast document repositories.

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

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

Acknowledgments

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

References

[1] T. Berners-Lee, J. Hendler and O. Lassila, "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, 2001, http://www.sciam.com/article.cfm?id=the-semantic-web
[2] Biogateway. http://www.semantic-systems-biology.org/home
[3] Y. Cao and Y. Li, "An intelligent fuzzy-based recommendation system for consumer electronic products". *Expert Systems with Applications*, vol. 33 (1), 2007, pp. 230-240.
[4] C. W. Cleverdon, J. Mills and E. M. Keen, "Factors Determining the Performance of Indexing Systems, vol. 2, Test Results". Cranfield: ASLIB Cranfield Project, 1966.
[5] J. Ferber, *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*, New York: Addison-Wesley Longman, 1999.
[6] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing". *International Journal of Human-Computer Studies*, vol. 43 (5-6), 1995, pp. 907-928.
[7] N. Guarino, "Formal ontology and information systems". In N. Guarino, (Ed.), *Formal Ontology in Information Systems*, Amsterdam :IOS Press, 1998, pp. 3-17.
[8] J. Hendler, "Agents and the Semantic Web". *IEEE Intelligent Systems*, March-April, 2001, pp. 30-37.
[9] F. Herrera, E. Herrera-Viedma and J. L. Verdegay, "Direct Approach Processes in Group Decision Making using Linguistic OWA operators". *Fuzzy Sets and Systems*, vol. 79 (2), 1996, pp. 175-190.
[10] F. Herrera and L. Martinez, "A 2-tuple fuzzy linguistic representation model for computing with words". *IEEE Transactions on Fuzzy Systems,* vol. 8 (6), 2000, pp. 746-752.
[11] E. Herrera-Viedma, E. Peis, J. M. Morales-del-Castillo and K. Anaya, "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*, Berlin: Springer Verlag, 2007, pp. 647-666.
[12] A. Isaac and E. Summers (Eds.), *SKOS Simple Knowledge Organization System Primer*, 2008. http://www.w3.org/TR/skos-primer/.
[13] D. L. McGuinness and F. van Harmelen (Eds.), *OWL Web Ontology Language Overview*, 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/
[14] National Cancer Institute Thesaurus. http://www.mindswap.org/2003/CancerOntology/
[15] M. Nottingham, *The Atom Syndication Format*, 2005. http://atomenabled.org/developers/syndication/atom-format-spec.php
[16] R. Oldakowsky and C. Byzer, *SemMF: A framework for calculating semantic similarity of objects represented as RDF graphs,* 2005. http://www.corporate-semantic-web.de/pub/SemMF_ISWC2005.pdf.
[17] C. L. Palmer, L. C. Teffeau and C. M. Pirmann, "Scholarly information practices in the online environment: Themes from the literature and implications for library service development". Report commissioned by OCLC Research, 2009. http://www.oclc.org/programs/publications/reports/2009-02.pdf
[18] PubMed. http://www.ncbi.nlm.nih.gov/pubmed/
[19] *RSS 2.0 at Harvard Law*, 2004. http://cyber.law.harvard.edu/rss/rssVersionHistory.html
[20] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Analysis of recommendation algorithms for e-commerce". In A. Jhingran, J.M. Mason and D. Tygar, (Eds.), *Proc. of ACM E-Commerce 2000 conference*, New York: ACM, 2000, pp. 158-167.
[21] A. I. Schein, A. Popescul and L. H. Ungar, "Methods and metrics for cold-start recommendations". In K. Jarvelin, M. Beaulieu, R. Baeza-Yates and S. H. Myaeng, (Eds), *Proc. of the 25'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, New York: ACM, 2002, pp. 253-260.
[22] B. Shapira, U. Hanani, A. Raveh and P. Shoval, "Information filtering: A new two-phase model using stereotypic user profiling", *Journal of Intelligent Information Systems*, vol. 8, 1997, pp. 155–165.
[23] L. A. Zadeh, "The concept of a linguistic variable and its applications to approximate reasoning". *Information Sciences*, vol. 8(1), 1975, pp. 199-249; vol. 8 (2), 1975, pp. 301-357; vol. 9 (3), 1975, pp. 43-80, 1975.

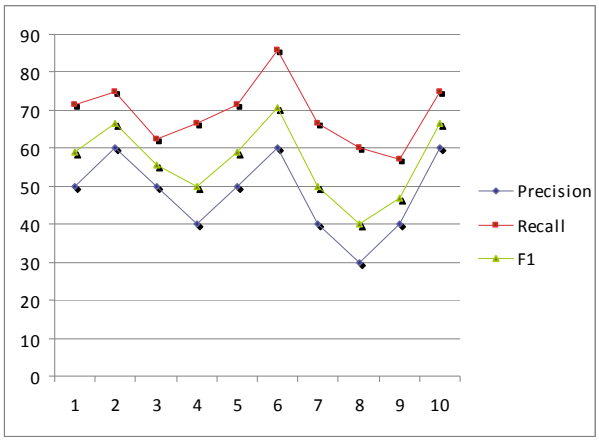

Figure 1. Precision, recall and F1

TABLE I. EXPERIMENTAL DATA

	Nrs	Nrn	Nis	Nr	Ns
User 1	5	2	5	7	10
User 2	6	2	4	8	10
User 3	5	3	5	8	10
User 4	4	2	6	6	10
User 5	5	2	5	7	10
User 6	6	1	4	7	10
User 7	4	2	6	6	10
User 8	3	2	7	5	10
User 9	4	3	6	7	10
User10	6	2	4	8	10

TABLE II. DETAILED EXPERIMENTAL OUTCOMES

%	User1	User2	User3	User4	User5	User6	User7	User8	User9	User10	**Aver.**
P	50.00	60.00	50.00	60.00	50.00	40.00	40.00	30.00	40.00	60.00	48.00
R	71.43	75.00	62.50	85.71	71.43	66.67	66.67	60.00	57.14	75.00	69.15
F1	58.82	66.67	55.56	70.59	58.82	50.00	50.00	40.00	47.05	66.67	56.42

Author Index

Mirbel, Isabelle..... 567
Mirea, Ana-Maria..... 215
Miura, Takao..... 283
Mladenic, Dunja..... 507
Molina, José M. 171
Morales-del-Castillo, J. M. 433
Moran, Stuart..... 327
Morgado, Fernando..... 129
Mullen, Tracy..... 607
Murata, Tsuyoshi..... 5
Nabuco, Olga..... 579
Nagata, Masaaki..... 100
Nagata, Naomi..... 219
Nakata, Keiichi..... 327
Nardini, Elena..... 501
Nasraoui, Olfa..... 91
Nassiri, Nasser..... 618
Navarro-Arribas, Guillermo..... 155
Navrat, Pavol..... 117
Nguyen, Giang-Son..... 466
Nica, Mihai..... 77
Ning, Wang..... 373, 381
Nishihori, Yuri..... 223
Nitta, Katsumi..... 357
Novotný, Róbert..... 121
Nozawa, Takayuki..... 9
Nunes, Sérgio..... 515
Oga, S. 349
Okabe, Masayuki..... 30
Okamoto, Toshio..... 219, 231
Olivas, Jose A. 159
Oliveira, Eugénio..... 515
Oliveira, Felipe F. 571
Oliver, Helen..... 50
Olivieri, Francesco..... 587
Omicini, Andrea..... 501
Orgun, Mehmet A. 474
Orii, Yuki..... 9
Papadakis, Ioannis..... 96
Park, GunWoo..... 445
Peis, Eduardo..... 433
Peischl, Bernhard..... 77
Peng, Zhang..... 381
Pes, Barbara..... 575
Petrone, Giovanna..... 42
Ping'an, Li..... 377
Pisano, Antonio..... 591
Poesio, Massimo..... 519
Popescu, Elvira..... 239
Popova, Anguelina..... 227
Porcel, C. 179
Porcel, Carlos..... 433
Pouliquen, Bruno..... 519, 523
Preda, Mircea Cezar..... 215
Psaila, Giuseppe..... 125, 163
Pudota, Nirmal..... 409
Qiuyan, Zhong..... 385
Qu, Weiguang..... 275
Quincey, Ed de..... 50
Raibulet, Claudia..... 361, 563
Ramos-Corchado, Félix F. 497
Ribaudo, Marina..... 207
Rizzo, Francesca..... 563
Rodrigues, Marcos..... 579
Rodríguez, R. M. 187
Romero, Elizabeth..... 91
Romero, Francisco P. 159
Ronchi, Stefania..... 125
Rotolo, Antonino..... 488
Rubens, Neil..... 231
Saez, Arturo..... 595
Sánchez-Pi, Nayat..... 171
Santos, Raphael O. 571
Santucci, Valentino..... 26
Sarmento, Luis..... 515
Sato, Haruhiko..... 223
Sattar, Abdul..... 474
Schadschneider, Andreas..... 583
Schmid, Wolfgang..... 77
Schroeder, Michael..... 50