2009 Ninth International Conference on Intelligent Systems Design and Applications

A web-based fuzzy linguistic tool to filter information in a biomedical domain

José Manuel Morales-del-Castillo
Dpt. Library and Information Science, University of Granada, Spain, josemdc@ugr.es

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

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

Abstract

In Biomedical Sciences is necessary the development of new services capable of satisfying specific information needs. In this paper we present a filtering 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

1. Introduction

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 [18]), 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 permit to improve the access to the 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 [6], the application of information filtering techniques [23], and the development of knowledge-based applications using Semantic Web technologies (such as the Biogateway Portal [3] or the *National Cancer Institute Thesaurus* [15]).

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 [25]. 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. An operational example of the performance of the system 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. Theoretical basis

The system here proposed is based on a previous multi-agent model defined by Herrera-Viedma et al. [13], 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

Manuscript received May 30, 2009 (date on which paper was submitted for review). Corresponding author: José M. Morales-del-Castillo (e-mail: josemdc@ugr.es).
Digital Object Identifier inserted by IEEE

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, for instance, RSS (*Really Simple Syndication* or *RDF Site Summary*) in any of its multiple versions [20].

The system is developed by the application of fuzzy linguistic modeling techniques and Semantic Web technologies to improve *user-agent* and *agent-agent* interaction, and to 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 system.

2.1. Semantic Web

The Semantic Web [2] 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: i) semantic tagging of resources, so that information can be understood both by humans and computers, and ii) the development of intelligent agents [10] 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 [1], 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 [8][9] 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. Fuzzy linguistic modeling approaches

Fuzzy linguistic modelling [25] 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 (7 or 9 tags):

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

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 $(a_i, b_i, \alpha_i, \beta_i)$ (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*: $s_i \geq s_j \quad if \quad i \geq j$.

$2.-$ *Negation operator*: $Neg(s_i) = s_j$, *with* $j = T - i$.

$3.-$ *Maximization operator*: $MAX(s_i, s_j) = s_i$ *if* $s_i \geq s_j$.

$4.-$ *Minimization operator*: $MIN(s_i, s_j) = s_i$ *if* $s_i \leq s_j$.

Besides, it is necessary to define aggregation operators, such as the *Linguistic Ordered Weighted Averaging* (LOWA) operator [11], 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 [12]. 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 $\beta \notin \{0, ..., g\}$ as a result of a symbolic aggregation of linguistic information [10], 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*" [12] which represents the difference between the information expressed by β and the nearest linguistic label $s_i \in S$.

2.3. 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 [19]), a set of RSS feeds containing the items to be recommended, a user profile repository and a test thesaurus in SKOS [14] format, that has been developed taking as a model the *National Cancer Institute Thesaurus* [16].

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 [15]. Furthermore, 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 [7] proposed by Willmott et al. [24], which is based on XML syntax and RDF/OWL as content language.

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 [17] which takes into account the position of the concepts to be matched in a taxonomic tree. Once defined the similarity between preferences and topic terms, 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 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 $P_i = \{\alpha_1, \ldots, \alpha_N\}$ 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:

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

- *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*: $S' x S \rightarrow S$:

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

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

where, (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) $\beta$is a bonus value which rewards or penalize the weights of the preferences. It is defined as β=*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 [11]. It 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.

3. 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 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,α_1, α_2
- 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 [17]. Let α_1 be the concept "*Vitamin E*" with a depth of 2 in the thesaurus of the system and β_2 the concept "*Suramin Sodium*" with a depth of 3 (being 6 the maximum depth of the thesaurus). The closest common parent (*ccp*) of both concepts is "*Angiogenesis Inhibitor*", which depth is 0 by default. As a result, the distance between α_1 and β_2 is *d* (α_1, β_2)= 0.83.

In this 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. ω_1=0.83) and for α_2 the label "*Medium*" (i.e. ω_2= 0.5). Besides, if the number of preferences and subjects is respectively N=2 and M=3, then the 3 maximum similarities are chosen to calculate the relevance value (in this case, let's suppose Sim ($\alpha_{1,}$ β_3)=0.88, Sim ($\alpha_{2,}$ β_1)=0.84, and Sim ($\alpha_{2,}$ 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 modeling 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 [10] the resulting aggregated label is the following: k= MIN{6,3 + round (0.4*(4-3))}=3 → 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 β=*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*".

4. Evaluation of the system

We have set up an experiment to evaluate the content-based module of the system in terms of precision [4] and recall [5] (since the collaborative recommendation module is not fully implemented yet and suffers from *cold start problem* [22]). These two measures (together with the F1 measure [21]) 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 [19], 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 which of them they consider as relevant. With these starting premises the experiment was carried out and the results are shown in table 1:

	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
Nrs	5	6	5	4	5	6	4	3	4	6
Nrn	2	2	3	2	2	1	2	2	3	2
Nis	5	4	5	6	5	4	6	7	6	4
Nr	7	8	8	6	7	7	6	5	7	8
Ns	10	10	10	10	10	10	10	10	10	10

Table 1. Experimental data.

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).

%	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	**Aver**
P	50.0	60.0	50.0	60.0	50.0	40.0	40.0	30	40.0	60.0	48.0
R	71.4	75.0	62.5	85.7	71.4	66.6	66.6	60	57.1	75.0	69.1
F1	58.8	66.6	55.5	70.5	58.8	50.0	50.0	40	47.	66.6	56.4

Table 2. Detailed experimental outcomes

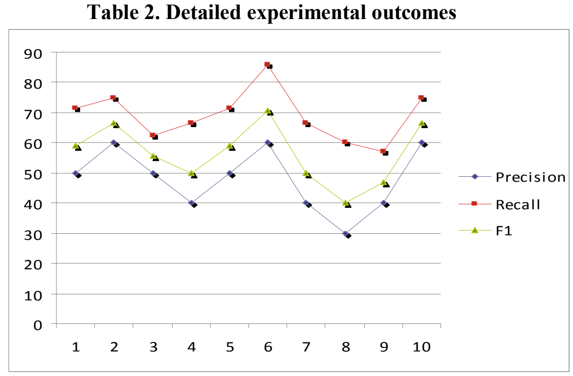

Figure 1. Precision, recall and F1

5. 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 techniques 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.

6. Acknowledgment

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

7. References

[1] D. Beckett (Ed.), *RDF/XML Syntax Specification*, 2004. http://www.w3.org/TR/rdf-syntax-grammar/

[2] 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

[3] Biogateway. http://www.semantic-systems-biology.org/home

[4] 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.

[5] Cleverdon, C.W., J. Mills and E. M. Keen, *Factors Determining the Performance of Indexing Systems, vol. 2, Test Results*. Cranfield: ASLIB Cranfield Project, 1966.

[6] Ferber, J., *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*, New York: Addison-Wesley Longman, 1999.

[7] Foundation for Intelligent Agents (FIPA). http://www.fipa.org

[8] 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.

[9] N. Guarino, "Formal ontology and information systems". In N. Guarino, (Ed.), *Formal Ontology in Information Systems*, Amsterdam: IOS Press, 1998, pp.3-17.

[10] J. Hendler, "Agents and the Semantic Web". *IEEE Intelligent Systems*, March-April, 2001, pp. 30-37.

[11] 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.

[12] F. Herrera and L. Martinez, "A 2-tuple fuzzy linguistic representation model for computing with words". *IEEE Trans. on Fuzzy Systems,* vol. 8 (6), 2000, pp. 746-752.

[13] 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 Liu, J., D. Ruan and G. Zhang, (Eds.), *E-Service intelligence: methodologies, technologies and applications*, Berlin: Springer Verlag, 2007, pp. 647-666.

[14] Isaac and E. Summers (Eds.), *SKOS Simple Knowledge Organization System Primer*, 2008. http://www.w3.org/TR/skos-primer/.

[15] D. L. McGuinness and F. van Harmelen (Eds.), *OWL Web Ontology Language Overview*, 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/

[16] National Cancer Institute Thesaurus. http://www.mindswap.org/2003/CancerOntology/

[17] 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.

[18] 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

[19] PubMed. http://www.ncbi.nlm.nih.gov/pubmed/

[20] *RSS 2.0 at Harvard Law*, 2004. http://cyber.law.harvard.edu/rss/rssVersionHistory.html

[21] 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.

[22] 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.

[23] B. Shapira, U. Hanani, A. Raveh and P. Shoval, "Information filtering: A new two-phase model using stereotypic user profiling", *J. of Intelligent Information Systems*, vol. 8, 1997, pp. 155–165.

[24] S. Willmott, F.O.F. Peña, C. Merida-Campos, I. Constantinescu, J. Dale and D. Cabanillas. "Adapting agent communication languages for semantic Web service inter-communication". In Skowron, A, et.al. (Eds.) *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence 2005*. New York: IEEE Society, 2005, pp. 405 – 408.

[25] 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.