Semantic Information Management for Pervasive Computing

Vassileios Tsetsos*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
b.tsetsos@di.uoa.gr

Abstract. In recent years, knowledge and semantics management has been applied to many areas of Informatics and Telecommunications. The main reason is that such management enables the creation of more flexible and usable computing environments, with higher added value for the users. In this thesis various different cases of semantics information management are studied, that refer to different aspects of modern and future computing paradigms. A special focus is put on context information management, when it is described through knowledge representation techniques. Specifically, in the first part of the thesis several methods for semantic service discovery are presented. Moreover, an evaluation framework for service discovery engines is proposed, which satisfies the special requirements of this discovery process. This framework, as shown by the experimental evaluation performed, is appropriate for evaluating such engines, since it is fully compatible with their special characteristics. In the second part the combination of context-awareness and information dissemination in autonomic computing environments is explored. The main goal for such combination is to be able to achieve collaborative context-awareness. The main contribution of this thesis is the design and evaluation of a efficient scheme for collaborative context-awareness. Finally, in the third part, a framework for personalized services is presented along with two case studies: semantic location services and personalized interactive TV. The proposed framework relies on rules and ontologies for providing advanced services to TV subscribers. In this part several issues of architectural and technological nature are addressed that are closely relevant to the implementation of such services.

Keywords: Pervasive Computing, Semantic Information Management, Context-Awareness, Mobile Computing, Semantic Services, Service Personalization

1 Introduction

The main objective of this thesis is to study ways to develop intelligent computing environments and advanced services. The techniques presented in the following sections are related to various concepts which are briefly described as follows:

* Dissertation Advisor: Stathes Hadjiefthymiades, Assistant Professor

- "Pervasive Computing": the term was first mentioned by Mark Weiser back to 1991. It refers to a new computing paradigm where computers, embedded in the physical user environment, are cooperating in a distributed way in order to provide an advanced user experience. Some key characteristics of this paradigm are a) system adaptation to changes in the user/application environment, b) user-centered and personalized applications, c) intelligent services.
- "Knowledge management and Semantic Web": semantic information is formally structured metadata that clearly define concepts and entities (through relationships, constraints, properties etc). Management of semantic information (or knowledge) refers to its representation, persistence, and update processes. A formal, yet practical, way of semantics representation is ontology. According to R. Studer (1998) "Ontology is a formal, explicit specification of a shared conceptualization". In practise, the most popular means for creating an ontology is the Semantic Web technologies (e.g., Web Ontology Language, Resource Description Framework Schema)
- "Context awareness": it refers to a system's ability to sense and react accordingly to changes in its operational parameters. A system may sense the context changes through sensors or other means. It can even rely on other systems to update it upon context changes, a technique called collaborative context awareness that was studied in one of the thesis' topics. Another important aspect of context awareness is its context modeling. Several context models have been proposed in the literature. In this thesis we explore knowledge representation methods for describing context.

Our thesis is that all the aforementioned aspects of computing, and especially the semantic technologies, form the basis for creating advanced and intelligent systems with the following main features:

- Distributed computation in heterogeneous environments.
- Collaborative computation
- Intelligent service behaviour and application personalization

1.1 Organization and contribution of the doctoral dissertation

The thesis is divided in three parts that study different aspects of semantic information management in pervasive computing environments. Special emphasis is put on investigating issues related to the features mentioned in the previous section.

The first part of the thesis deals with service discovery and, particularly, its evaluation. Since service discovery is a core functionality of pervasive computing, we aimed to design a framework for assessing the effectiveness of service discovery systems. Two topics were investigated:

a) finding appropriate metrics for evaluating service discovery systems

b) creating a methodology for evaluating discovery systems with graded relevance scale and without relevance judgments from experts.

The contribution of our research can be summarized to:

a) two metrics that are used in generalized information retrieval systems were proposed. These are suitable for evaluating graded discovery results and are more accurate than the standard Precision and Recall metrics.

b) an evaluation method that combines various techniques from the information retrieval field was proposed. It does not require relevance judgments from experts since it automatically generates pseudo-relevance judgments. Then it exploits the aforementioned metrics in order to evaluate and compare the systems.

The second part of the thesis refers to collaborative context awareness through efficient message exchange in nomadic (ad hoc) environments (e.g., vehicular networks). In such environments, there is often the need for intelligent context aware services. An efficient way to achieve this is through knowledge representation and reasoning techniques. However, in such environments not all nodes do have sensors and their resources (e.g., energy, storage) are usually limited. Hence, the nodes have to collaborate efficiently in order to deliver the expected functionality. The contribution of the thesis in this area is a publish/subscribe scheme where the nodes exchange sensor-originating data [1]. The data model adheres to a context ontology, thus enabling support of knowledge management processes. The scheme makes minimal assumptions for the underlying network infrastructure and nodes and tries to be sensitive in terms of event detection (i.e., context changes). Its performance is shown to be superior than that of other schemes which can be applied to nomadic environments (i.e., periodic polling).

In the third part of the thesis we study issues related to personalization of context-aware services and applications, with the aid of semantics. Specifically, a generic framework is presented for developing such applications. Basic elements of such framework are models that represent all the involved entities, with more important being the user model. Modelling of such applications is based on ontologies and the actual personalization process is implemented through rules and the respective reasoning engines. The proposed framework was validated in two application domains: a) semantic location based services (navigation and other services in users with disabilities), and b) personalization of multimedia services in interactive TV environments (content and service provisioning based on the program semantics and the viewer's profile). In the area of semantic location based services, a user model is combined with a location (spatial) model so that the system can search for accessible paths, taking into consideration the user's disabilities or preferences. Usage of such extensive knowledge representation techniques in this application domain is considered as one of the thesis' key contributions [2][3].

The same applies to the second application domain, too [4] (i.e., semantic interactive TV). There have been proposed several similar systems in the literature [5], however none of them uses formal semantics technologies in order to provide the desired personalization. Such declarative way to affect the provisioning of value added services is, in our view, a key factor for their adoption and success.

2 Evaluation of Semantic Web Service Discovery

Service discovery and selection are central topics in distributed systems research. Semantic Web Services (SWS) are an evolution of Web Services that are based on metadata and allow for more expressive description of service capabilities, used both for service advertisements and requests. Such metadata is represented through well-

known knowledge representation tools, like ontologies and rules, implemented with Semantic Web technologies. The SWS paradigm involves, apart from the aforementioned description facilities, more sophisticated (mainly logic-based) matchmaking algorithms [6]. Typical information that the SWS discovery (SWSD) systems exploit are attributes such as: service Inputs, Outputs, Preconditions and Effects (a.k.a. IOPE attributes). However, no matter on which elements of a service description the matchmaking algorithm is applied to, the most important problem in matchmaking is that it is unrealistic to expect relevant advertisements and requests to be in perfect match. The problem is aggravated if we take into account that the service request may not fully capture the requestor's intention. The concept of the "Degree of Match" (DoM), a kind of a relevance scale, was introduced for dealing with this problem [6].

Similarly to other retrieval systems, such as Web search engines, SWS discovery systems should be evaluated in terms of performance and retrieval effectiveness. Many researchers have already undertaken performance assessment efforts for measuring retrieval times and the scalability of the available tools (an extensive review of SWSD engines can be found in [6]). However, to the best of our knowledge, only a few researchers have performed such experimental evaluations. There are several reasons for this situation, with the most important outlined below:

1. Lack of established evaluation metrics. The typical metrics used in Information Retrieval are not directly applicable and they do not fully take into account the semantics of the degrees of match. Moreover, the existing metrics do not assume service rankings with ties (aka weak or partial rankings) in their majority. Such rankings are typically returned by the discovery engines.

2. Lack of (sufficiently large) test collections. Current service test collections, either for plain WS or SWS have a small number of services and an even smaller number of queries and relevance judgments [7].

3. Incomplete relevance judgments. In general, SWSD should be evaluated with methods that assume incomplete or totally missing relevance judgments. This is a very important and realistic assumption in open environments like the Web.

2.1 Background on Evaluation of Service Discovery Processes

The entities involved in a service discovery process are the service advertisements (Si) published in a service registry, the service request R posed by the user, and the matchmaking engine that is responsible for the actual service discovery. In essence, the matchmaking engine assigns a Degree of Match e(R, Si) to every service advertisement Si. These values determine the ranking of the final advertisements for a specific request R. In order to evaluate the matchmaking engine effectiveness some expert mappings r(R, Si) (i.e., relevance judgments between R and each Si) should be available/pre-specified. Hence, the vectors r and e are defined as:

$$r: Q \times S \rightarrow W, \quad e: Q \times S \rightarrow W$$

where Q is the set of all possible service requests, S the set of service advertisements and W the set of values denoting the degree of relevance (for r) or degree of match (for e) between a request from Q and a service from S. Both r and e may assume various types of values: Boolean (W={0,1}), real numbers (W=[0,1]), fuzzy terms (W={"irrelevant", "relevant", …}), etc. Given these informal definitions, the

evaluation of a matchmaking engine is the determination of how closely vector e (delivered by the engine) approximates vector r (specified by domain experts).

A Boolean evaluation scheme is the traditional scheme used in the relevant literature. In this case, standard measures such as precision and recall are used for measuring the system performance. However it has some considerable pitfalls:

- the graded results of matchmaking algorithm execution are transformed to Boolean values, and, thus, the matchmaking and service semantics is ignored,
- such transformation involves the definition of a threshold. The assignment of an optimal value to this evaluation parameter is not a trivial task and there are no formal and commonly agreed ways it can be done, and
- the Boolean relevance assessments are too coarse-grained and do not always reflect the real intention of the domain expert

Given these shortcomings, the Boolean evaluation scheme cannot accurately assess how close the discovered services are to the actual relevant services. A solution to these problems would be to use an evaluation scheme based on graded relevance. However, this implies that apart from the existence of graded relevance judgments, appropriate metrics are in place. In the following sections we review some metrics that have been proposed in the literature and propose necessary adjustments.

2.2 Related Work

Most of the initial approaches rely on the Boolean evaluation scheme. To our knowledge, the first attempt to apply the concept of graded relevance to the evaluation of service discovery is described in [8]. This work was followed by [9] that proposed new metrics and methods for evaluating SWSD. However, even if the authors exploit graded relevance and reach some very interesting conclusions, they overlook the fact that the rankings returned by the SWSD tools are partial. Moreover, the fact that their evaluation relies on manually created relevance judgments constitutes a limiting factor as was the case for [8].

Besides the standard retrieval evaluation metrics (Precision, Recall, F-measure etc.) other metrics have been proposed in the literature. A very popular metric is Average Precision (AveP) over all relevant items for a query/request. Other approaches for catering for incomplete relevance judgments are RankEff, Ap_all, InducedAP, SubCollectionAP and InferredAP. In [10] Kekäläinen and Järvelin introduced the concept of gain. Every level of the relevance scale holds a gain value which indicates the gain that the user receives from finding a service belonging to this relevance level. They also proposed Cumulated Gain (cg) at rank r which depicts the total gain that the user receives by exploring the resulted ranking till rank r. However, Cumulated Gain does not penalize late retrieval of relevant services. Hence, the authors also added a discount factor in the calculation which decreases the gain of services as the rank increases, resulting in Discounted Cumulated Gain (dcg). To be able to compare various DCG curves from different discovery engines they proposed the use of normalised dcg (nDCG). nDCG is computed by dividing the DCG value of the result ranking with the dcg of an optimal ranking, called idcg (ideal dcg).

Based on the concept of gain, various metrics have been proposed, which can be expressed in terms of Cumulated Gain, like Q-Measure [11] and Average Weighted

Precision (AWP) [9]. Sakai in [11] proposed a new metric for graded relevance called Q-Measure which integrates AveP and AWP. Küster and König-Ries in [9] proposed Average Weighted Discounted Precision (AWDP) to be used for the evaluation of SWS discovery. However, all the metrics described so far are capable of evaluating only full rankings of services, or items in general, i.e. rankings without ties. McSherry and Najork [12] proposed a method to extend metrics for full rankings to partial rankings. In their work they extended Precision, Recall, F1-measure, Reciprocal Rank, Average Precision and nDCG (we will call nDCGp' its version applied to condensed lists). From all of them only the last one supports graded relevance.

Finally, a basic issue in evaluation is the creation of a test collection with relevance judgments. This task is performed manually, as some experts must judge every item in the collection with respect to every query. The solution that is proposed in [13] is the use of pooling. This method of pooling can be used when the set of items is a finite set that does not change frequently. But this is not the case for dynamic environments like the World Wide Web (WWW), where both the set of items and queries are under frequent change and it would be useful to evaluate systems without the need of manually created relevance judgments. In that direction, various techniques have been proposed, that can be classified in two large categories, those who automatically create relevance judgments from the rankings returned [13] and those who evaluate systems without relevance judgments.

2.3 Proposed Metrics for evaluation of SWS discovery

In this section we investigate and propose some evaluation metrics that demonstrate the desired characteristics for SWSD evaluation, as reported in the previous section. The first proposal is based on generalized versions of Precision and Recall. The other metrics are adaptations of metrics already proposed by other researchers.

2.3.1 Generalized Metrics

In order to deal with the problems identified above, we can assume that a service discovery system is a generalized retrieval system. In [8] we proposed such an evaluation scheme and performed a preliminary evaluation. The main idea is that the domain experts assess the relevance of specific services against a given request through fuzzy linguistic terms. In order to be able to compare the degrees of match used by the engines with the corresponding expert relevance assessments we need to express them in a similar form. Moreover, new metrics are required in this case. The proposed measures are generalizations of the recall and precision measures, calculated from the two rankings of relevance assessments: those delivered by the engine and those performed by domain experts in a way similar to the Boolean case.

2.3.2 Metrics for Evaluating Partial Rankings

From the literature survey it became apparent that apart from the nDCGp' metric [12], Q-Measure and AWDP would also capable of evaluating SWSD systems, if they supported partial rankings. Hence, we provided some extensions of these metrics

towards this direction. The first extension was the definition of AWDP for partial rankings. The second one was the Q-measure for partial rankings.

2.4 Automatic Generation of Relevance Judgments

As already mentioned, it is important to be able to assess the effectiveness of SWSD systems without manually created relevance judgments, since it is very difficult to obtain them from domain experts. In this section we study two social voting methods for creating pseudo-relevance judgments. The first of these methods is based on the Borda Count and the second is the Condorcet method. Condorcet method seems to have better properties for the task at hand. A brief description of the method follows.
According to the Condorcet method, the voters (matchmaking engines) rank the candidates (services) in an order of preference. However, ties and incomplete rankings are allowed. The final ranking is calculated based on the pair-wise number of wins of each candidate (i.e., service).

Firstly, the method builds a comparison matrix CM of size n x n, where n is the cardinality of the set of all services returned by the engines. The CM[i,j] element denotes the "wins" of the service *i* over the service *j*. When a service has not been returned by an engine, then it is assumed that has been defeated by all other ranked services returned by the engine. Next, for each pair of services we find the final ranking. We compare pair-wise all services and we add one point for each pair-wise win, lose and tie in the respective columns. The rules for ordering the services are:

- Winner is the service with the most wins.
- If two services have the same number of wins, then winner is the service with the less defeats.
- If the number of wins equals the number of defeats, then the services are ranked equally.

Our final goal, however, is to align this ranking to the relevance scale we use, so that the final relevance judgments are generated. In order to do this we use the following formula:

$$f_i = \frac{m - rank_i}{m}$$

where m is the number of discrete relevance levels in the final ranking, and $rank_i$ is the position of the i^{th} element in the ranking.

2.5 Experimental Evaluation

In order to assess the applicability of the metrics we performed several experiments under variable settings. The main objective of the experiments was to explore the behavior of the metrics and techniques used.

Setup

In the evaluation experiment we involved the following matchmaking engines [6]:

- OWLS – MX, which provides five different matchmaking algorithms.

- OWLS-SLR, which performs logic-based DL reasoning and also implements two distance metrics.
- TUB OWL-S Matcher, which performs matchmaking based on DL subsumption over many service parameters.

Something very interesting is the type of ranking returned by each engine. Table 1 shows a categorization of the aforementioned engines according to the type of ranking they return (as assumed for the experiments).

Table 1. SWSD engine categorization based on the type of ranking returned

SWSD Engine	Ranking
OWLS-MX (M0)	Partial
OWLS-MX (M1-M4, M1mx2-M4mx2)	Full/Partial
OWLS-SLR (UC, ED)	Full/Partial
TUB	Partial

In order to measure the effectiveness of the engines that returned full rankings we used the nDCG', Q'-measure and AWDP' metrics (quotes mean that they are applied to condensed lists). We also used these latter metrics in order to measure the effectiveness of engines returning partial rankings so as to assess the error introduced by their misuse. In order to present aggregate results for the generalised precision we defined the Average Generalised Precision (Average P_G or APG) which is the sum of all P_G values for all relevant services divided by the number of relevant services.

For the experiments we relied on the TC3 service collection. The main characteristics of this collection are that it includes a relatively large number of services and service requests and the corresponding relevance judgments as decided by "human experts". Apart from TC3, we also used the automatically-generated relevance judgments generated with the Condorcet method. Both TC3 and pseudo-relevance judgments had the same relevance scale shown in Table 2.

Table 2. Relevance scale for TC3 and pseudo-relevance judgments

Rank	Range of rank	Gain
Highly Relevant	(0.75, 1]	3
Relevant	[0.5, 0.75)	2
Potentially Relevant	[0.25, 0.5)	1
Irrelevant	[0, 0.25)	0

Experiments and Results

Firstly, we study the correlation between the automatically generated relevance judgments (pseudo-relevance judgments) and the TC3 judgments.

The pseudo-relevance judgments were produced based on the Condorcet method. The Kendall's tau-b correlation between the results was computed between the condensed lists of pseudo-relevance judgments and the TC3 mappings. We used the Kendall's tau-b correlation coefficient because it takes into consideration ties among

the ranked items. For each query, only the services that had been discovered by all engines were used for calculating the correlations (even if they had been judged as totally irrelevant).

Finally, in order to check the "rationality" of pseudo-relevance judgments, we manually created the following variants of the TC3 relevance judgments:

- *TC3+30* : the value of 30% of the relevance judgments (selected randomly) was upgraded to the next higher level in the relevance scale (e.g., from "Potentially Relevant" to "Relevant"),
- *TC3-30* : the value of 30% of the relevance judgments (selected randomly) was downgraded to the next lower level in the relevance scale,
- *TC3r30* : the value of 30% of the relevance judgments for each request were modified randomly, either upgraded or downgraded by one level.

Each variant is supposed to "represent" a different expert behavior.

The results are shown in Fig. 1 (one can observe that the Condorcet judgments have the highest correlation with the original TC3 mappings).

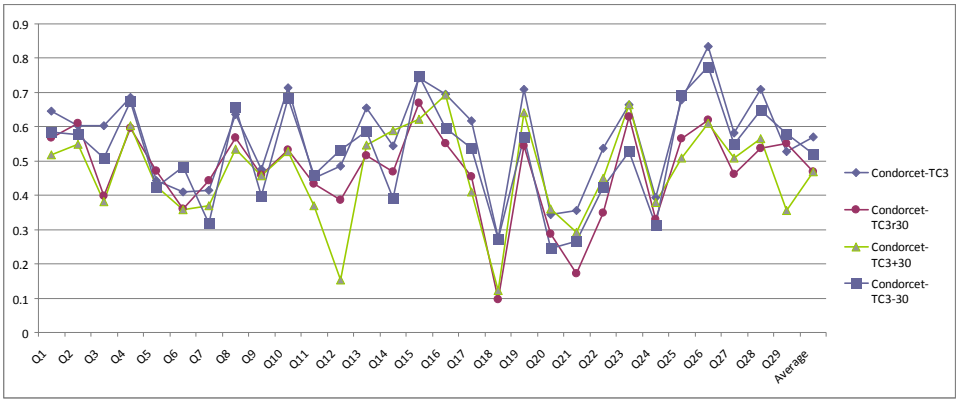

Fig. 1. Kendall's tau-b correlation: Condorcet and variants of TC3

Next, we show how much the SWSD evaluation depends on the selection of the correct metrics. Some aggregate results for better comparison are presented in Fig. 2 (for partial rankings/metrics and TC3 relevance judgments) and in Fig. 3 (same as previous but using the Condorcet relevance judgments). In all these figures, crossed lines denote differences between the rankings of the engines by the metrics. The correlation between the results of the various metrics is presented in Table 3.

Some observations that can be extracted from these figures are:

1. All metrics, except for the (average) nDCGp' agree on the relative order of the engines' effectiveness (Fig. 2). nDCGp' does not take into account the total number of relevant services (for which there exist relevance judgments), hence engines like the M2mx2 that return very few, but relevant, services per request, are ranked at a very high position. Hence, we can safely conclude that this metric is not very appropriate for the domain of SWSD evaluation.

2. The AGP values are quite similar to AWDP(V) and Q-Measure(V) values, despite the fact that relies on a completely different approach, i.e., it is not based on the concept of *gain* (Fig. 2 and Table 3).

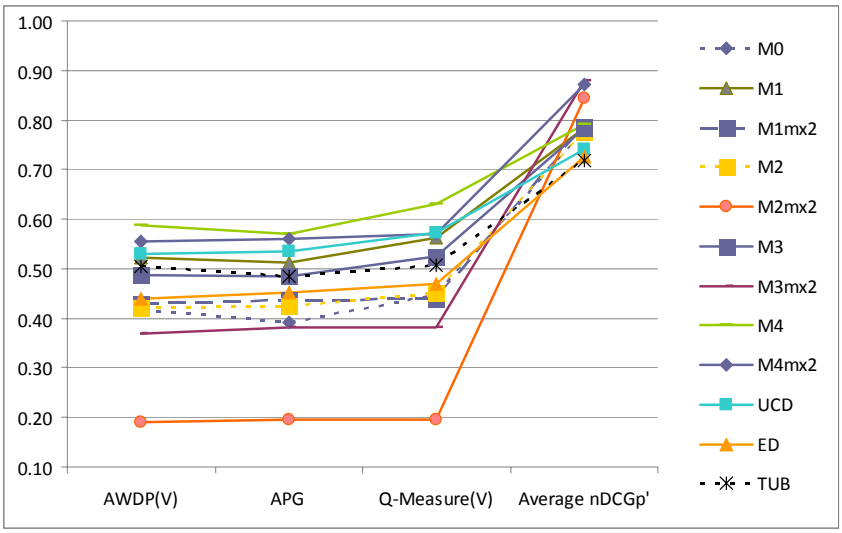

Fig. 2. Comparison of metrics and engines for partial service rankings (TC3)

Table 3. Pair-wise correlation of metrics (left:TC3, right: Condorcet)

		Kendall's tau
AWDP(V)	APG	0.96
AWDP(V)	Q-Measure(V)	0.89
AWDP(V)	nDCGp'	-0.18
APG	Q-Measure(V)	0.92
APG	nDCGp'	-0.14
Q-Measure(V)	nDCGp'	-0.23

		Kendall's tau
AWDP(V)	APG	0.73
AWDP(V)	Q-Measure(V)	0.90
AWDP(V)	nDCGp'	0.29
APG	Q-Measure(V)	0.63
APG	nDCGp'	0.00
Q-Measure(V)	nDCGp'	0.40

The measurements show that the SWSD engines that participated in the creation of the pseudo-relevance judgments are favored in the experiments that use such judgments. However, we can always decide which are the best and worst services for a specific request. An observation is that the nDCGp' metric has the worst performance for the pseudo-judgments while the APG the best one.

As shown in Table 3, the APG metric has a high correlation coefficient with the other two "reliable" metrics, AWDP(V) and Q-measure(V). This fact, in combination with the observation of the previous subsection that it is less affected by the use of pseudo-judgments, constitutes it a rather promising metric. Another observation is that the absolute APG values are not affected significantly by the set of relevance judgments used (Fig. 2 and 3). Engines like M0 and TUB are slightly rewarded (since they participated in the creation of pseudo-judgments), but in general most of the engines have similar value ranges in the two experiments. In contrast, the AWDP(V) and Q-measure(V) metrics increase significantly the effectiveness values of the engines in Fig. 3. This is another reason why the he generalized metrics seem more appropriate when the pseudo-judgments are used.

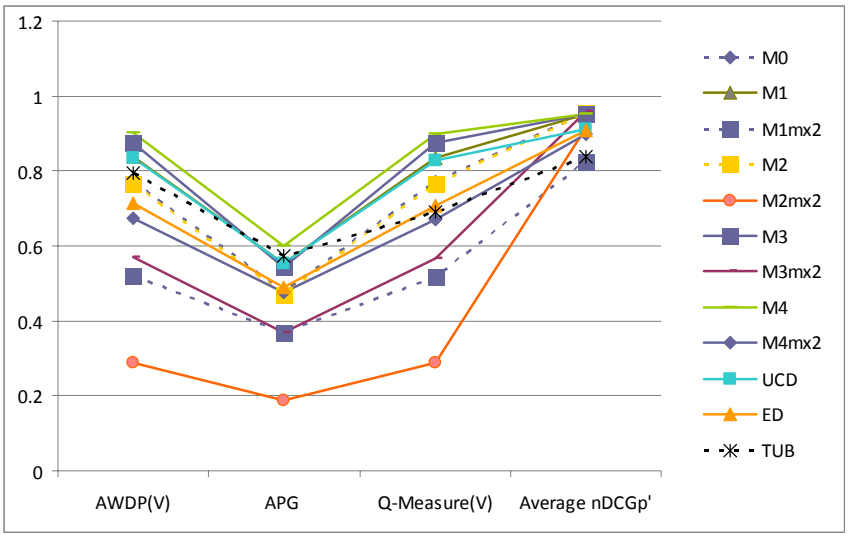

Fig. 3. Comparison of metrics and engines for partial service rankings (Condorcet)

Finally, as shown in the Fig. 2 the most effective service matchmaking engine is M4 (all metrics agree on that). Quite close performance have the UCD and M1 engines. In any case, M2mx2 engine has the worst service discovery performance. These results agree with the findings of the creators of the respective engines, as stated in the respective papers.

3 Conclusions

In the present thesis, several aspects of using semantic information management in building advanced applications and systems were studied. Some general conclusions follow:

Semantic Web Services

Semantic Web Services (SWS) are among the most usable and interesting results in the research field of the Semantic Web. Several researchers have dealt with SWS discovery and composition. However, there is a considerable lack of methods and tools for evaluating their efforts. In this work we proposed some methods in this respect, and we assessed their applicability experimentally. More research is necessary for establishing standard means of evaluation in this domain.

Collaborative context awareness

Context and situation awareness is an important aspect of modern and future computing systems. In this work we tried to increase the overall level of contact awareness in a distributed environment, through a collaborative asynchronous scheme. The results were quite promising, since the performance of the system improved without decreasing the capability of the nodes to detect changes in their environment.

Semantic personalization

One of the most widely known forms of application intelligence is personalization. We proposed a reference framework that can facilitate the semantic description of entities and adaptation rules. The framework was validated in two application domains. In conclusion, semantic information management, where possible, can solve significant design problems of personalized systems.

References

1. Tsetsos, V. and Hadjiefthymiades, S. "An Innovative Architecture for Context Foraging", Eighth International ACM Workshop on Data Engineering for Wireless and Mobile Access (MobiDE, in conjunction with SIGMOD/PODS 2009), Providence, Rhode Island, (2009)
2. Tsetsos, V., Anagnostopoulos, C., Kikiras, P., and Hadjiefthymiades, S., "Semantically enriched navigation for indoor environments," International Journal of Web and Grid Services, vol. 2, no. 4, Inderscience Publishers, pp. 473--478, (2006)
3. Papataxiarhis, V., Riga, V., Nomikos, V., Sekkas, O., Kolomvatsos, K. Tsetsos, V. Papageorgas, P. Xouris, V., Vourakis, S., Hadjiefthymiades, S., and Kouroupetroglou, G., "MNISIKLIS: Indoor LBS for All", *5th International Symposium on LBS & TeleCartography (LBS 2008)*, Salzburg, Austria, November, (2008).
4. Tsetsos, V., Papadimitriou, A., Anagnostopoulos, C. and Hadjiefthymiades, S. "Integrating Interactive TV Services and the Web through Semantics", *to appear in International Journal On Semantic Web and Information Systems, SI on "Semantic Media Adaptation & Personalization"*, IGI-Global, (2010)
5. Fernandez, B., Pazos Arias, Y., Lopez Nores, J.J., Gil Solla, A., & Ramos Cabrer, M. AVATAR: An Improved Solution for Personalized TV based on Semantic Inference, IEEE Transactions on Consumer Electronics. 52(1), pp. 223—231, (2006).
6. Tsetsos V., Anagnostopoulos C., and Hadjiefthymiades S.,"Semantic Web Service Discovery: Methods, Algorithms and Tools", chapter in *"Semantic Web Services: Theory, Tools and Applications"* (Ed. Dr. Jorge Cardoso), IDEA Group Publishing, (2007)
7. Fan, J. and Kambhampati S. A Snapshot of Public Web Services, SIGMOD Record, 34(1), pp. 24--32, 2005).
8. Tsetsos, V., Anagnostopoulos, C., Hadjiefthymiades, S. "On the evaluation of Semantic Web Service matchmaking systems", 4th IEEE European Conference on Web Services (ECOWS), Zurich, Switzerland, (2006)
9. Küster, U., and König-Ries, B. "Evaluating Semantic Web Service Matchmaking Effectiveness Based on Graded Relevance," in ISWC '08, Karlsruhe, Germany, (2008).
10. Kekäläinen, K., and Jaana, J. "IR evaluation methods for highly relevant documents," in SIGIR '00, pp. 41—48, (2000)
11. Tetsuya Sakai, "New performance metrics based on multigrade relevance: Their application to question answering," in NTCIR '04, Tokyo, Japan, (2004).
12. McSherry, F., and Najork, M. "Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores," in Lecture Notes in Computer Science. Berlin Heidelberg: Springer-Verlag, (2008).
13. Nicholas C., & Cahan P. Soboroff I., Ranking retrieval systems without relevance judgments.: In Proceedings of the 24th ACM SIGIR conference, (2001).