User Interests Modeling Based on Multi-source Personal Information Fusion and Semantic Reasoning

Yunfei Ma[1], Yi Zeng[1], Xu Ren[1], and Ning Zhong[1,2]

[1] International WIC Institute, Beijing University of Technology, Beijing, China
mafly008@emails.bjut.edu.cn, yizeng@bjut.edu.cn
[2] Department of Life Science and Informatics, Maebashi Institute of Technology, Maebashi-City, Japan
zhong@maebashi-it.ac.jp

Abstract. User interests are usually distributed in different systems on the Web. Traditional user interest modeling methods are not designed for integrating and analyzing interests from multiple sources, hence, they are not very effective for obtaining comparatively complete description of user interests in the distributed environment. In addition, previous studies concentrate on the text level analysis of user interests, while semantic relationships among interests are not fully investigated. This might cause incomplete and incorrect understanding of the discovered interests, especially when interests are from multiple sources. In this paper, we propose an approach of user interest modeling based on multi-source personal information fusion and semantic reasoning. We give different fusion strategies for interest data from multiple sources. Further more, we investigate the semantic relationship between users' explicit interests and implicit interests by reasoning through concept granularity. Semantic relatedness among interests are also briefly illustrated for information fusion. Illustrative examples based on multiple sources on the Web (e.g. microblog system Twitter, social network sites Facebook and LinkedIn, personal homepage, etc.) show that proposed approach is potentially effective.

1 Introduction

User interests have shown their increasing importance in driving the development of personalized Web services and user-centric applications. Existing studies on analyzing user interests focus on browsing behaviors (such as duration) and browsing contents (such as viewed Web pages) [1,2]. These methods can only get users' previous interests. Meanwhile, the obtained user interests are limited by the contents of the viewed Web pages. Another direction for obtaining user interests is to ask users to have direct inputs or provide feedbacks (such as evaluating the resources, or adding tags) [3], but sometimes users do not have positive attitudes to take part in these activities. In addition, many users cannot provide a relatively complete list of his/her interests since user interests are usually distributed in different environments.

N. Zhong et al. (Eds.): AMT 2011, LNCS 6890, pp. 195–205, 2011.

In this paper, we focus on finding user interests directly from user generated contents. User interests might be distributed in different sources on the decentralized Web platform (e.g. microblog, social network site(SNS), homepage, etc.). Hence, we propose to integrate user interests from these sources. The idea and methods of information fusion is brought to obtain user interests from these heterogeneous sources. Considering the characteristics of these different sources, we propose a weighted fusion approach for multi-source user interests modeling.

User interests are divided into explicit interests and implicit interests [4]. Explicit interests are defined as user interests that are explicitly stated by users in some way. Implicit interests are inferred ones from explicit interests [5]. Text mining is a commonly used approach for inferring user interests, nevertheless, it is hard for this approach to infer accurate hierarchical relationships among interests. In order to solve this problem, in this paper, we utilize semantic reasoning with domain ontology to infer implicit interests from users' explicit ones. User interests are characterized as domain concepts, and implicit interests are obtained by inference with superclass and subclass relations. By using this approach, the context of the explicit interests can be acquired.

Personal information fusion on user interests helps to integrate, analyze, and understand user interests distributed in the decentralized Web platform. Semantic reasoning helps to infer implicit interests and produce contextual understanding of the discovered interests. By using these two approaches, one can get relatively complete understanding of a specific user's interests, and produce more personalized services for them.

2 The Framework of Multi-source Personal Information Fusion for User Interests Modeling

2.1 The Workflow of Multi-source User Interests Modeling

User interests are distributed in different Web-based systems and platforms in the Web age. Each user may hold several accounts, while each of them are related to some unique user interests related data. In order to make the user interests modeling process more accurate and complete, a workflow need to be designed and several major steps need to be considered. Here we briefly discuss each step.

***S*tep 1.** Information sources selection and user data extraction. In this step, different sources which are related to user interests need to be selected. After selection of resources, with user authorization, user related data need to be extracted for user interests discovery.

***S*tep 2.** Single-source user interests discovery. In this step, user interests are discovered by keywords extraction and analysis from each single source. Different keywords analysis methods can be applied, such as cumulative interests statistics, retained interests, interests durations, etc [6].

***S*tep 3.** Multi-source interests fusion. In this step, discovered interests from Step 2 are integrated together by information fusion strategies to produce a relatively complete ranked list of the specified user.

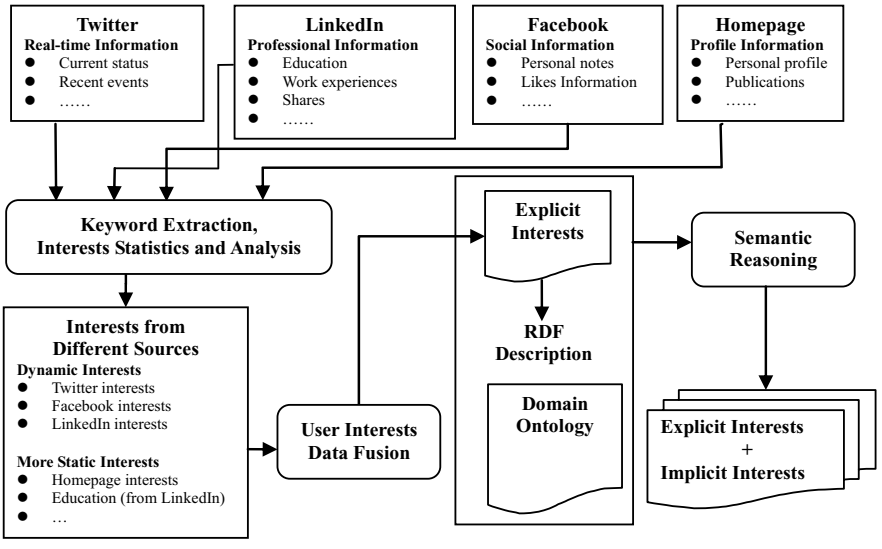

Fig. 1. The Workflow of Multi-source User Interests Modeling

***S*tep 4.** User interests and their ontology description. In this step, user interests need to be described by knowledge representation languages. User interests ontology need to be produced based on these descriptions.

***S*tep 5.** Semantic reasoning on existing interests. In this step, reasoning techniques are applied to produce implicit interests based on existing ones. In this way, one can understand user interests in contexts and the implicit interests also help to refine the user interests ontology.

Figure 1 gives an overview of these steps. In the next section, we will discuss each steps in details. Novel methods that are designed for each step will be proposed. Meanwhile, concrete illustrative examples and discussions will be given in the context of real data from the Web.

2.2 Multi-source Information Extraction for User Interests Modeling

As is mentioned above, we plan to integrate user's personal information from different sources. In the current Web environment, we choose to get user information from various social network sites (SNS), homepages, professional networks, etc. Different types of SNS record different kinds of information about user. For example, Twitter records one's real-time status and most recent interested events in 140 characters, while Facebook stores relatively longer notes, likes and shared resources. Homepages and professional networks contain user's education information, work experiences, publications and long term interests (e.g. LinkedIn), etc.

Hence, we choose Twitter, Facebook, LinkedIn and homepage for multi-source personal information extraction. These selections cover most personal data that

are publicly accessible. As shown in Figure 1. Open APIs of these Web platforms make sure the accessibility of related user data. (Twitter, Facebook, and LinkedIn provide their own API for downloading user related data, while homepage information can be crawled through Google Web API). We should emphasize that the information sources we selected can be adjusted according to the actual situation of a specific user. For instance, if a specific user may have no Twitter account but have a Sina Weibo (a Chinese microblog similar to Twitter) account, then we should also add this site as a source.

Generally speaking, user interests can be represented by keywords. Hence, keyword extraction is essential in user interests modeling. An interest may need single-word term or multi-word term to represent. If we only consider single-word term, space between words can be used for segmenting interests. If we consider multi-word term, more complex term extraction algorithms and tools need to be applied. In our study, we use AlchemyAPI for keywords extraction[1].

2.3 Dynamic and Static User Interests

User related information from multiple sources can be roughly divided into two types, dynamic information and static information. The dynamic information refer to the information with created or update time (e.g. Tweets, Facebook notes, LinkedIn information). While the static information refer to those with no time tags (e.g. professional interests, affiliation, and education information from one's homepage). These two different types of information should be separately considered, since we can extract different types of interests from them. Dynamic information contains dynamic interests with tagged time slots, and static information contains static interests. In most cases, static interests can be treated as long-term interests, otherwise, users will not state them on relatively static sources such as homepage.

Dynamic and static interests have different usages in specific applications. Although they sometimes have overlaps (e.g. "Semantic Web" can be found both on the author Frank van Harmelen's Twitter and on his homepage), they should be treated separately. In our study, dynamic interests are analyzed statistically and ranked list of interests are with values, while static interests are organized as an independent interests set, and they will not be ranked together with dynamic interests. In personalized Web applications, static interests, such as affiliation and location, serve as additional contextual information of the specified user, while dynamic interests serve as implicit constraints for user activities.

3 User Interests Fusion Strategies

3.1 Decision Level Fusion for User Interests Modeling

Although microblog, social network sites, professional network sites, and homepage are different types of Web-based systems, they can be considered as various

[1] AlchemyAPI is a product of Orchestr8, LLC, a provider of semantic tagging and text mining. Term extraction by AlchemyAPI is based on statistical natural language processing and machine learning (*http://www.alchemyapi.com/*)

sensors that provide user related data from different perspectives. Hence, the integration of user interests from multiple Web-based systems can be considered as multi-sensor information fusion.

Based on information fusion theory, fusion strategies can be divided into three types, namely, data level fusion, feature level fusion and decision level fusion [7]. Twitter, Facebook and LinkedIn, etc. can be considered as different types of sensors. Considering that they capture user interests from different aspects, data level fusion may not be appropriate. In this paper, we consider decision level fusion on user interests, since we want to compare the user interests sets generated from different single sources and the ones produced by different fusion strategies. In decision level fusion of user interests, several list of ranked interests are generated based on each individual sources, and the fusion processes are executed based on a certain kind of fusion strategy.

3.2 A Weighted Fusion Approach for Multi-source User Interests Modeling

The generation of ranked interests lists can be based on various interests ranking strategies [6]. In this paper, we select the cumulative interest function to rank interests and indicate users' preference among them. In cumulative interest function, each interests are assigned with a value of their word frequency during a specified time slot [6]. After the steps of keywords extraction and ranking by cumulative interest values, we get the ranked lists of interests data from different sources.

In order to get more complete and holist understanding of the specific user's interests. We need to integrate these interest lists from multiple sources. In this paper, we propose a weighted fusion approach for multi-source user interests modeling. The fusion equation can be represented as follows:

$$I(i) = \sum_{n=1}^{m} w_n \cdot I(i)_n \tag{1}$$

where i represents a specific interest, $I(i)_n$ represents the interest value of i from the nth single source. w_n denotes the weight of the nth source, which can be determined according to different specific strategies. $I(i)$ denotes the user interest value we get after the fusion. Here we provide two concrete strategies. One is the average fusion strategy and another is time-sensitive fusion strategy. In average fusion, $w_n = 1/n$, which means every source is assigned the same weight. In time-sensitive fusion, w_n is determined by the following equations:

$$\begin{aligned} &w_1 : w_2 : ... : w_n = f_1 : f_2 : ... : f_n \\ &w_1 + w_2 + ... + w_n = 1 \end{aligned} \tag{2}$$

where f_n is the information update frequency of the nth source (the average number of relevant messages released by the specific user per day). The equation shows that the fusion weights are positive relevant to the information update rate. Since most homepages do not maintain update time information, in our experiment, we only consider the interest fusion from Twitter, Facebook notes and LinkedIn.

3.3 Experimental Results and Analysis

Here we select "Frank Van Harmelen" (an Artificial Intelligence researcher) as an example to make a comparative study on different fusion strategies. For time-sensitive fusion, we get his information update rates in Twitter ($f_1 = 2.5$), Facebook ($f_2 = 0.2$), and Linkedin ($f_3 = 0.0004$) accordingly. Hence, $w_1 = 0.9258$, $w_2 = 0.0741$, and $w_3 = 0.0001$. The time-sensitive interest fusion function is represented as:

$$I(i) = 0.9258 \cdot I(i)_1 + 0.0741 \cdot I(i)_2 + 0.0001 \cdot I(i)_3 \tag{3}$$

We choose the top 10 interests keywords in a time interval to illustrate the fusion process (Interests are ranked in decreasing order on the values). Results

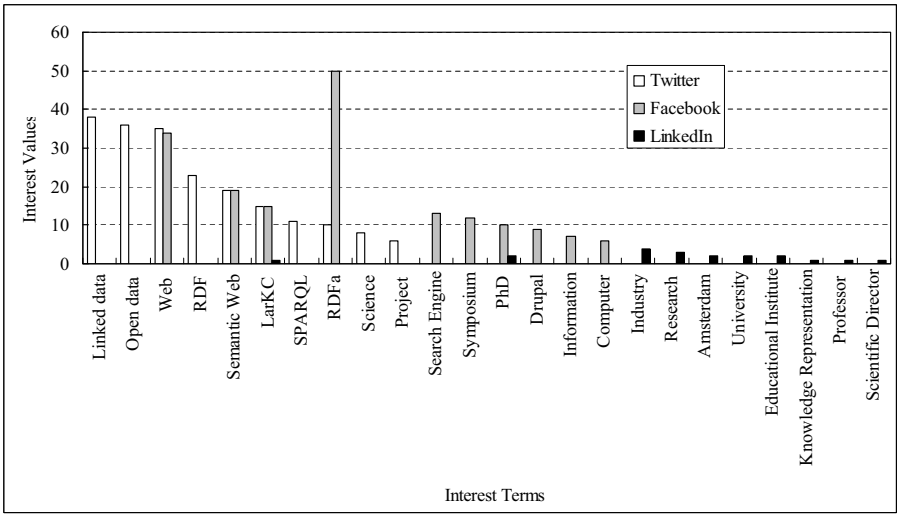

Fig. 2. A Comparative Study of Interests Ranking of Single sources

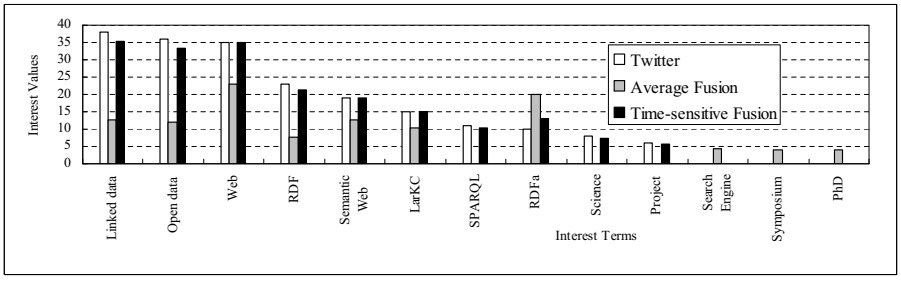

Fig. 3. A Comparative Study of Interests Ranking from Single sources and Multi-source Fusion

from single sources are shown in Figure 2. We can observe that interests from different sources may have overlaps (e.g. "Semantic Web", "LarKC", and "RDFa" from Twitter and Facebook), while they may also have many diversities (e.g. there are only one in common between the interests from Twitter and the ones from LinkedIn, namely "LarKC").

Results based on the average fusion strategy and time-sensitive fusion strategy are shown as in Figure 3. The result list of average fusion strategy contains 7 interests from Twitter, 7 interests from Facebook and 2 interests from LinkedIn. The result list of time-sensitive fusion strategy contains the same interests terms with Twitter, and the interests values are all very relevant to the ones from Twitter. Except for overlapped interests terms, the fusion list does not contain the interests that only appear in the lists of Facebook and Linkedin.

As observed from Figure 3, the result of time-sensitive fusion is highly relevant to the ranked list from Twitter, since Frank updates his Twitter much more frequent than his Facebook notes and LinkedIn. On the other hand, the sequence of the same interests in Twitter and in time-sensitive fusion list are not exactly the same ("Web" and "Open Data" have swapped their positions, so did "RDFa" and "SPARQL", which are mainly caused by interests from Facebook). Hence, if we want to get fusion results that are more real-time, we should apply the time-sensitive fusion strategy. If time is not a very important factor, and each sources need to be realized, the average fusion strategy will be better. In addition, one can develop their own strategies to decide on the weights for these multiple sources.

4 Semantic Reasoning to Infer Implicit Interests

User interests are not isolated texts, they might be related to each other from the semantics perspective. In order to have deeper utilization of user interests, they need to be represented by knowledge representation languages. In addition, semantic reasoning can be applied to the represented user interests so that implicit interests can be discovered.

4.1 Representation of User Interests in RDF

When representing user interests from multiple sources, static interests and dynamic interests need to be represented separately. For dynamic interests, they are organized as a ranking list in this paper, and each interest is assigned with a value. We adopt the e-FOAF:interest vocabulary [2] to represent user's dynamic interests [8]. Here we give a fragment of the author Frank van Harmelen's interests profile based on time-sensitive fusion.

```
<foaf:Person rdf:about="http://www.cs.vu.nl/~frankh/">
  <foaf:name>Frank van Harmelen</foaf:name>
    <e-foaf:interest>
```

[2] E-foaf:interest Vocabulary Specification *http://wiki.larkc.eu/e-foaf:interest*

```
      <rdf:Description rdf:about="http://www.wici-lab.org/wici/
             wiki/index.php/Web">
      <dc:title>Web</dc:title>
      <e-foaf:cumulative_interest_value rdf:parseType="Resource">
        <rdf:value rdf:datatype="&xsd;number">34.922</rdf:value>
      </e-foaf:cumulative_interest_value>
      </rdf:Description>
    </e-foaf:interest>
    ...
</foaf:Person>
```

For static interests, they are organized as an interests set. In the representation of these interests, value property of each interest is ignored. One can put representation of dynamic interests and static interests into the same RDF file. They can be distinguished by whether they have value descriptions. Alternatively, they also can be represented into two separate files so that they can be selectively loaded for different needs.

```
<foaf:Person rdf:about="http://www.cs.vu.nl/~frankh/">
  <foaf:name>Frank van Harmelen</foaf:name>
    <foaf:topic_interest>AI Department</foaf:topic_interest>
    <foaf:topic_interest>Semantic Web</foaf:topic_interest>
    <foaf:topic_interest>The Netherlands</foaf:topic_interest>
    ...
</foaf:Person>
```

By using FOAF vocabularies, the upper fragment is an illustrative example of Frank van Harmelen's static interests. As shown in the example, only interests terms are provided, and they do not have a strict order.

4.2 Finding Implicit Interests by Reasoning on Interests Hierarchy

Reasoning can help to find implicit knowledge based on existing facts. In our study, in order to have a contextual and more complete understanding of user interests, we need a further step of reasoning to expand the explicit user interests list by the implicit ones.

Many domain knowledge can be organized as a hierarchical ontology, and each term of this domain can be distributed in different levels which are with different granularities [9,10]. Most interest terms can be considered to be from a certain domain, hence, hierarchical ontology can help to get a specific interest's context in different levels of granularities.

From explicit interests, we can approximately predict user's main research field. Figure 4 presents a fragment of a domain ontology for "Artificial Intelligence" [3]. We represent this ontology in RDF, with the same method introduced in [11], as shown in the following.

[3] Here we do not discuss whether this ontology is well designed, we only show how to obtain implicit interests by using this ontology

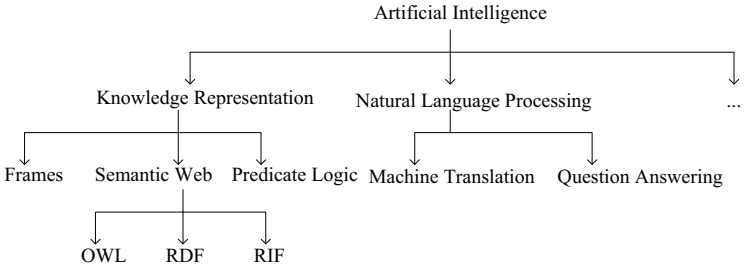

Fig. 4. A Fragment of An Artificial Intelligence Ontology

```
<rdfs: Class rdf: ID="Semantic Web">
  <rdfs:subClassOf rdf: resource="Artificial Intelligence"/>
</rdfs:Class>
<rdfs: Class rdf: ID="RDF">
  <rdfs:subClassOf rdf: resource="Semantic Web"/>
</rdfs:Class>
```

The process of finding implicit interests based on domain ontology can be described as follows:

***S*tep 1.** Locate the specified explicit interest on the domain ontology.

***S*tep 2.** Perform inference by certain kind of reasoning (such as reasoning with rdfs:subClassOf or rdfs:superClassOf relation).

***S*tep 3.** Extend interests list by interests acquired through reasoning.

If we want to get interests with coarser levels of granularity than the explicit interests, we try to reason out the superclass of the existing ones, and if we want to get interests with finer levels of granularity compared to the explicit ones, we try to reason out the subclass of the explicit ones. Thus a hierarchical context of explicit interests can be acquired and the original interest list can be expanded. Since domain terms are usually organized on several levels, it might not be practically effective if all levels are considered for expansion. We suggest it would be better to expand the interest list one level coarser or finer than the explicit interests.

For example, "Semantic Web" is a interest in Frank van Harmelen's time-sensitive interest fusion list, as shown in Figure 3. We locate this keyword on the hierarchical ontology in Figure 4. By applying reasoning rule for finding superclass, we can conclude that Frank is interested in "Knowledge Representation". This fact is in the interests list of LinkedIn, although it is not explicit in the time-sensitive interest fusion list, it can be inferred from this list. We also can get the fact that Frank is generally interested in Artificial Intelligence.

5 Conclusion and Future Work

In this paper, we presented a framework of multi-source personal interests fusion. We described the workflow of the proposed method and illustrated the different phases of the approach. Two steps are of vital importance, namely, interest fusion from multiple sources and semantic reasoning to extend the interest list. For interest fusion, we proposed a weighted fusion function together with two concrete strategies (i.e. average fusion and time-sensitive fusion). Illustrative examples are provided based on the data from multiple sources such as Twitter, Facebook, LinkedIn, etc. We should claim that for some users, their personal information on these platforms are not public, and the proposed approach is only effective for users who are not mind to share their data.

In the future, we will continue to improve the proposed approach. In this paper, we mainly focus on the fusion of dynamic interests. In future studies, we are going to investigate on how to integrate dynamic interests and static interests, and meanwhile to realize the difference between them. Secondly, except for the proposed average fusion strategy and time-sensitive fusion strategy, we are going to work on other possibilities for interests fusion from multiple sources. In this paper, we only introduced reasoning with hierarchical relations. Other possibilities need to be considered for producing implicit interests, such as extending interest lists by reasoning with semantic similarity [5,12].

Acknowledgement. This study is supported by Beijing Postdoctoral Research Foundation (2011ZZ-18), China Postdoctoral Science Foundation (20110490255), and the Large Knowledge Collider (LarKC) Project (FP7-215535) under the European Union 7th framework program.

References

1. Liang, T.P., Lai, H.J.: Discovering user interests from web browsing behavior: An application to internet news services. In: Proceedings of the 35th Annual Hawai'I International Conference on Systems Sciences, pp. 2718–2727. IEEE Press, Los Alamitos (2002)
2. Seo, Y.W., Zhang, B.T.: Learning user's preferences by analyzing web browsing behaviors. Artificial Intelligence 15(6), 381–387 (2001)
3. Carmagnola, F., Cena, F., Cortassa, O., Gena, C., Torre, I.: Towards a tag-based user model: How can user model benefit from tags? In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS (LNAI), vol. 4511, pp. 445–449. Springer, Heidelberg (2007)
4. Kim, H.R., Chan, P.K.: Learning implicit user interest hierarchy for context in personalization. Applied Intelligence 28(2), 153–166 (2008)
5. Zeng, Y.: Unifying Knowledge Retrieval and Reasoning on Large Scale Scientific Literatures. PhD thesis, Beijing University of Technology (2010)
6. Zeng, Y., Zhou, E., Wang, Y., Ren, X., Qin, Y., Huang, Z., Zhong, N.: Research interests: Their dynamics, structures and applications in unifying search and reasoning. Journal of Intelligent Information Systems 37(1), 65–88 (2011)

7. Varshney, P.K.: Multisensor data fusion. Electronics & Communication Engineering Journal 9(6), 245–253 (1997)
8. Zeng, Y., Wang, Y., Huang, Z., Damljanovic, D., Zhong, N., Wang, C.: User interests: Definition, vocabulary, and utilization in unifying search and reasoning. In: An, A., Lingras, P., Petty, S., Huang, R. (eds.) AMT 2010. LNCS, vol. 6335, pp. 98–107. Springer, Heidelberg (2010)
9. Calegari, S., Ciucci, D.: Granular computing applied to ontologies. International Journal of Approximate Reasoning 51(4), 391–409 (2010)
10. Yao, Y.: A Unified Framework of Granular Computing. In: Handbook of Granular Computing, pp. 401–410. Wiley, Chichester (2008)
11. Zeng, Y., Zhong, N., Wang, Y., Qin, Y., Huang, Z., Zhou, H., Yao, Y., van Harmelen, F.: User-centric query refinement and processing using granularity based strategies. Knowledge and Information Systems 27(3), 419–450 (2011)
12. Wang, Y., Wang, C., Zeng, Y., Huang, Z., Momtchev, V., Andersson, B., Ren, X., Zhong, N.: Normalized medline distance and its utilization in context-aware life science literature search. Tsinghua Science and Technology 15(6), 709–715 (2010)