User Interests Driven Web Personalization Based on Multiple Social Networks

Yi Zeng
International WIC Institute
Beijing University of Technology
Beijing, China
yizeng@bjut.edu.cn

Ning Zhong
Department of Life Science and Informatics
Maebashi Institute of Technology
Maebashi-City, Japan
zhong@maebashi-it.ac.jp

Xu Ren, Yan Wang
International WIC Institute
Beijing University of Technology
Beijing, China
wang.yan@emails.bjut.edu.cn

ABSTRACT

User related data indicate user interests in a certain environment. In the context of massive data from the Web, if an application wants to provide more personalized service (e.g. search) for users, an investigation on user interests is needed. User interests are usually distributed in different sources. In order to provide a more comprehensive understanding, user related data from multiple sources need to be integrated together for deeper analysis. Web based social networks have become typical platforms for extracting user interests. In addition, there are various types of interests from these social networks. In this paper, we provide an algorithmic framework for retrieving semantic data based on user interests from multiple sources (such as multiple social networking sites). We design several algorithms to deal with interests based retrieval based on single and multiple types of interests. We utilize publication data from Semantic Web Dog Food (which can be considered as an academic collaboration based social network), and microblogging data from Twitter to validate our framework. The Active Academic Visit Recommendation Application (AAVRA) is developed as a concrete usecase to show the potential effectiveness of the proposed framework for user interests driven Web personalization based on multiple social networks.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval; H.3.5 [**Information Systems**]: On-line Information Services

General Terms

algorithm, human factors, experimentation

Keywords

interest analysis, Web personalization, search refinement

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WI&C'12 April 16, 2012 Lyon, France
Copyright 2012 ACM 978-1-4503-1189-2/12/04 ...$10.00.

1. INTRODUCTION

The adoption of personalization has been proved to be an efficient approach for finding useful information in massive data on the Web [1, 3, 4]. The advances of Web based social networks provide various data sources for user interests extraction. Users might express their interests in different social network sites. Hence, user interests related data are distributed in different social data sources. Methods for user interests integration and analysis are needed to understand user interests in the distributed environment. In [2], we provide a weighted function for integrating user interests from multiple data sources, and validate the proposed function by the data from various social networks (i.e. Facebook, Twitter, and LinkedIn). Nevertheless, the analysis of user interests is restricted to only one type of interests (namely, research interests), and the paper also did not investigate on how to use the analyzed user interests. In addition, in more complex scenarios, the types of user interests varies in different data sources, and the importance of different data sources might also be different, but they all might be related to a specific service (such as personalized search).

In this paper, we investigate on how to utilize interests data in personalized Web search from the algorithmic perspective. More importantly, we provide an algorithmic framework for interests based retrieval in the context of multiple interests types from multiple data sources (social networks). In order to validate the proposed algorithms, we developed the Active Academic Visit Recommendation Application (AAVRA) based on academic social networking data (from Semantic Web related publication network) and microblogging data from Twitter. (We restrict our processing data to RDF/OWL semantic data. Other types of data need to be transformed to RDF/OWL format so that they can be processed under the same framework and platform.)

2. AN ALGORITHMIC FRAMEWORK FOR INTERESTS BASED RETRIEVAL

Interests based retrieval method emphasizes that user interests is one of the most important heuristic factors for finding most relevant RDF triples to a specific user in the context of Web-scale knowledge bases [6, 7].

Interests can be divided into different types. Such as research interests, interests for visiting a place, interests for collaborating with someone on a paper, etc. In the simplest case, there is only one type of interest related relationship, such as someone "collaborate with" someone, or someone "visited" a place (Here "collaborate with" and "visited" are relationships that define one type of interest respectively).

In this section, we are going to provide an algorithmic framework for interests based retrieval in the context of single types of interests and multiple types of interests from multiple social networks.

2.1 Interests Based Retrieval with Single Interest Type

For single interest type, we firstly consider that a specific user only have one interest, and the selection process is only based on this specified interest.

Let i be a specific interest, and $f(t, i)$ be an evaluation function that represents the frequency of the interest i's appearance in the triple t (namely, $f(t, i)$ is a nonnegative integer) in the triple set T ($T = \{t \mid t =< s, p, o >\}$). Triples in T are ranked according the value of $f(t, i)$. Interests based retrieval strategy assumes the end user is interested in the triples that contain i, no matter which positions it appears in the RDF triple (Namely, the positions of subject, predicate, and object are all considered).

Let $R(t, i)$ be the ranked number of the triple t in the ranked triple set T' according to $f(t, i)$ (Namely, $R(t, i)$ is a positive integer). $R(t, i)$ is negative relevant to $f(t, i)$, namely, for two arbitrary triples $t, t' \in T'$, they satisfy that:

$$f(t, i) > f(t', i) \Longrightarrow R(t, i) < R(t', i).$$

Algorithm 1 (denoted as IRSI algorithm) illustrates the major steps for interests based retrieval with a single interest. The retrieval process is an iterative process. Each time, Top-K triples according to $f(t, i)$ are selected for processing. If there is still extra time, another Top-K triples will be selected from the rest of the triple set.

Algorithm 1: Interests-based **R**etrieval with a **S**ingle **I**nterest (IRSI)

```
Input: user interest $i$, original RDF triple set $T$
Output: a set of Top-$K$ triples
1 Begin
2 Calculate $f(t, i)$ for each $t \in T$, and obtain $R(t, i)$ according
  to $f(t, i)$;
3 Re-rank each $t$ according to $R(t, i)$ and generate an ordered
  triple set $T'$;
4 Select Top-$K$ triples from $T'$ as output and mark the
  selected triples for further processing;
5 If time allows and the user are not satisfied with the selected
  triples, output the selected triples from Step 4, and check
  whether unselected triples are available. If yes, goto Step 4
  to select new triples from $T'$, else, go to Step 6;
6 End
```

In most cases, a specific user may have multiple interests, which can be denoted by different interests terms. Each interest may be quantitatively evaluated. Hence, the selection of interesting triples is based on a weighted evaluation function.

Let i_m be the mth interest in the interests set I (i.e. $i_m \in I$), $F(t, I)$ be the weighted evaluation function for selecting RDF triples, and it is defined as:

$$F(t, I) = \sum_{m=1}^{n} \omega_m f(t, i_m), \tag{1}$$

where ω_m is the weight for the interest i_m according to a specific interest evaluation function. The interest evaluation function that assign a weight to a specific interest can be defined from different perspectives (e.g. cumulative interests, retained interests, interests longest duration, and interests cumulative duration function introduced in [6, 7] can be used). The original triple set T is re-ranked to an ordered triple set T' according to $F(t, I)$. The rank number of the specific triple t in the ranked triple set T' is represented as $R(t, I)$, and it is negative relevant to $F(t, I)$. Namely, for any two arbitrary triples $t, t' \in T'$, they satisfy that:

$$F(t, I) > F(t', I) \Longrightarrow R(t, I) < R(t', I).$$

Algorithm 2 (denoted as IRMI algorithm) focus on interests based retrieval with multiple interests.

Algorithm 2: Interests-based **R**etrieval with **M**ultiple **I**nterests (IRMI)

```
Input: original RDF triple set $T$
Output: a set of Top-$K$ triples
1 Begin
2 Select an interest evaluation function from [7] and acquire a
  set of interests $I$ as well as their values;
3 Calculate $f(t, i)$ for each $t \in T$ and $i \in I$;
4 Calculate $F(t, I)$ for each $t \in T$ according to Equation 1 (the
  weights of each interest is decided by the interest value
  calculated in Step 2), and obtain $R(t, I)$ according to $F(t, I)$;
5 Re-rank each $t$ according to $R(t, I)$ and generate an ordered
  triple set $T'$;
6 Select Top-$K$ triples from $T'$ as output and mark the
  selected triples for further processing;
7 If time allows and the user are not satisfied with the selected
  triples, output the selected triples from Step 6, and check
  whether unselected triples are available. If yes, goto Step 6
  to select new triples from $T'$, else, go to Step 8;
8 End
```

Theoretically, it would cost more processing time compared to only considering one interest, nevertheless, the selected triples by Algorithm 2 would be much more relevant to the user compared to the results from Algorithm 1. In [7], we provide concrete examples on how to utilize the proposed algorithm in personalized scientific literature search.

2.2 Interests Based Retrieval based on Multiple Data Sources

On the Web platform, a specific user's interests might be distributed in multiple data sources (such as various social networks). If the interests related triples from different sources only describe one type of user interests, then Algorithm 1 and Algorithm 2 can be used for ranking and selecting RDF triples. If they belong to different types of interests, they need to be organized as different sub triple sets and ranking among these triple sets need to be done before analysis on each type of interest.

Let $\mathcal{T}$ be a set of triples that contains different types of user interests. They are divided into different sub triple sets according to some constraints and an ordered set of sub triple sets (denoted as $\mathcal{T}'$) is generated (i.e. $\mathcal{T}' = \{T_1, T_2, ..., T_n\}$). The order of these sub triple sets is based on specific application scenario. Let T_x and T_y be two arbitrary triple sets that contain different types of interests. Their rank orders are assigned according to a specific scenario. For example, $R(T_x) < R(T_y)$. In this case, the retrieval process will be executed on T_x first, when all the triples in T_x are processed, then the retrieval process will go to T_y. Within T_x (or T_y), Algorithm 1 and Algorithm 2 are used for ranking of triples for Top-K triples selection. This process is formalized as Algorithm 3.

Algorithm 3: Interests-based **R**etrieval based on **M**ultiple **D**ata **S**ources (IRMDS)

```
  Input: triple set $\mathcal{T}$ with RDF triples from multiple sources
  Output: a set of Top-$K$ triples
1 Begin
2 Categorize the original RDF triple set $\mathcal{T}$ into several sub
  sets according to their types.
3 Rank the sub sets into an ordered triple sets
  $\mathcal{T}' = \{T_1, T_2, ..., T_n\}$;
4 Select an RDF triple set $T_x$ from the set of ordered RDF
  triple sets $\mathcal{T}'$ for further processing, and delete the selected
  one from $\mathcal{T}'$;
5 Apply Algorithm 1 and Algorithm 2 within the selected
  triple set;
6 If time allows, goto step 2;
7 End
```

The assignment of sub triple sets orders is usecase sensitive, and may differ from each other in different scenarios according to different requirements in various usecases. If the multiple data sources are various social networks, the order of triple sets can be based on different social relations from these social networks. A target data driven method for automatic data sources order generation is discussed in [5].

3. ACTIVE ACADEMIC VISIT RECOMMENDATION APPLICATION : A USECASE

In this scenario, we aim to provide an Active Academic Visit Recommendation Application ($AAVRA$ for short) for scientific researchers in an active way as long as their personal information can be acquired through (semantic) data on the Web. Users need to log in the system with their name or account, and specify which country or city they want to visit. Then, the recommendation system will extract their interests from multiple data sources to automatically build their interests profiles for recommendation.

In this scenario, the Semantic Web Dog Food (SWDF for short) dataset[1] is used. In addition, user related data (such as user profile information, persons whom the user follows, and tweets that the user wrote) are generated as an RDF triple set from Twitter in real-time.

User interests can be divided into many types. For SWDF data, author-publication related triples reflects specific authors' research interests and their interests in collaboration with other authors. For Twitter data, they reflect users' interests on certain topics and their interests in following someone, or commenting on, or retweeting their tweets. These types of interests are considered to be different from each other. Hence, this scenario is designed to be used to verify the interests based retrieval method based on multiple interests types from multiple data sources.

In $AAVRA$, we want to help users find interesting places for academic visit automatically and actively. Since the scenario is restricted to academic visit, we emphasize that if a user is interested in collaborating with someone from a place (namely, the explicit or potential collaborators are the user's interests), then he/she is interested in visiting the place. The recommendation is designed to be based on explicit and potential personal relationships. In our study, we generate the ranking of different interests types based on the types of social relations from various social networks.

Here we consider 5 levels of interests for a specific user in this scenario, and we have the following predicate denotations as constraints to formalize these levels:

- $SWDF(p)$: p is a person who is an author in the Semantic Web Dog Food dataset.
- $Coauthor_{SWDF}(p, u)$: p and u are two arbitrary persons who satisfy $SWDF(p) \wedge SWDF(u)$, and they are coauthors based on the record in the Semantic Web Dog Food dataset.
- $PCoauthor_{SWDF}(p, u)$: p and u satisfy $SWDF(p) \wedge SWDF(u) \wedge \neg Coauthor_{SWDF}(p, u)$.
- $Twitter(p)$: p is a person who owns a Twitter account.
- $TFing(u, p)$: p and u are two arbitrary persons who satisfy $Twitter(p) \wedge Twitter(u)$ and u is following p.
- $SIT(p, u, K)$: p and u satisfy $Twitter(p) \wedge Twitter(u)$, and they have at least K interests in common based on the analysis of their Tweets.

Recommendations in different levels of interests for $AAVRA$ can be decided by the following formula in Table 1:

Table 1: Interests Level Division for $AAVRA$

Level	Triple Set	Formula
1	T_1	$Coauthor_{SWDF}(p, u) \wedge TFing(u, p)$
2	T_2	$Coauthor_{SWDF}(p, u) \wedge \neg TFing(u, p)$
3	T_3	$TFing(u, p) \wedge PCoauthor_{SWDF}(p, u)$
4	T_4	$TFing(u, p) \wedge SIT(p, u, K) \wedge \neg SWDF(p)$
5	T_5	$TFing(u, p) \wedge \neg SIT(p, u, K) \neg SWDF(p)$

As can be observed in Table 1, different levels of interests correspond to different sub data sets generated according to different social relation based combinatorics strategies. Based on the context of this application, we assign that: $R(T_1) < R(T_2) < R(T_3) < R(T_4) < R(T_5)$. Hence, according to Algorithm 3 in Section 2.2, the recommendation will be executed over these levels of interests sequentially.

The whole system is developed as a workflow and run on the Large Knowledge Collider (LarKC) platform [2], with several plugins performing different functionalities. The workflow for the $AAVRA$ usecase is shown as follows:

$\boxed{\text{Twitter-Interests Identifier}} \rightarrow \boxed{\text{SWDF Identifier}} \rightarrow$
$\boxed{\text{Interests-based Selector}} \rightarrow \boxed{\text{Geo-Location reasoner}}$

Twitter-Interests Identifier is used to identify interesting contents from Twitter that are related to specific users (e.g. following, tweets, profiles). SWDF Identifier is used to identify relevant triples that are related to the specific user

[1] RDF/XML dumps of Semantic Web Dog Food data: <http://data.semanticweb.org/dumps/>

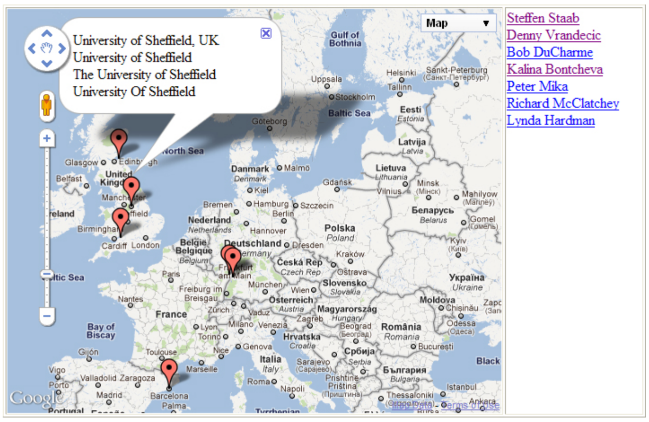
[2] LarKC: A massive semantic data processing platform for the Web <http://www.larkc.eu>.

(e.g. coauthors and affiliation related triples) from the Semantic Web Dog Food dataset. Interests-based Selector is used to select user interests related triples through different combinatorics analysis strategies introduced in Table 1. Geo-Location reasoner is used to find the location of the recommendation according to the affiliation of recommended candidates and mark recommendations on Google Maps.

Table 2 presents a partial example on different levels of interests for "Frank van Harmelen" in $AAVRA$ by using Algorithm 3. The results are provided to the end users levels by levels. The recommendation places for academic visit are based on these selected persons. In Table 2, The ratio of recommendation is acquired by $\frac{Number\ of\ recommended\ results}{Problem\ Space}$. The problem space is the total number of authors who satisfy $SWDF(p)$ or $TFing("Frank\ van\ Harmelen", p)$, namely 7131 persons. As has been indicated in this example, the number of possible recommendations can be dramatically reduced (only 0.883% of the persons in the problem space are recommended), and the recommended ones are highly relevant to the end user's background and requirement. Within a specific level, results can be ranked based on Algorithm 1.

Table 2: Different Levels of Interests for "Frank van Harmelen" in $AAVRA$

Interests Levels	Ratio of Recommendation	Sample Results
1	0.014%	Paul Groth
2	0.210%	Spyros Kotoulas(3), Jacopo Urbani(3), Eyal Oren(2), Henri Bal(2), Zharko Aleksovski(2), ...
3	0.154%	Kalina Bontcheva, Lynda Hardman, Steffen Staab, Denny Vrandecic, Ivan Herman, ...
4	0.505%	Stefano Bertolo, Dan Brickley, DERI Galway, Web Foundation, Ontotext AD ...



University of Sheffield, UK
University of Sheffield
The University of Sheffield
University Of Sheffield

Steffen Staab
Denny Vrandecic
Bob DuCharme
Kalina Bontcheva
Peter Mika
Richard McClatchey
Lynda Hardman

Figure 1: Academic Visit Recommendation for "Frank van Harmelen" to the U.K.

Figure 1 presents an example of academic visit recommendation to the U.K for the user "Frank van Harmelen". Some recommendations are made based on his potential collaborators (authors in the Semantic Web Dog Food, but still not collaborators) and at the same time serves as persons whom the user follows on Twitter (i.e. interests discussed in Level 2). The recommended places are: University of Sheffield (where Kalina Bontcheva is from), University of the West of England (where Richard McClatchey is from), etc.

4. CONCLUSIONS

In this paper, we have discussed how the interests from different social networks can be integrated and ranked. Social relations from different social networks are used for deriving and ranking different types of interests. The validation results of $AVVRA$ shows that a specific personalized application (more specifically personalized search and recommendation system) might require the support of interests analysis and ranking from multiple social networks, and the recommendation results are much more comprehensive compared to the use of only one data source. The proposed algorithmic framework serves as an applicable approach for interests based retrieval under the context of multiple social networks.

Acknowledgments

This study was supported by research grants from the Large Knowledge Collider (LarKC) Project (FP7-215535) under the European Union 7th framework program, the Young Scientists Fund of the National Natural Science Foundation of China (61100128), the China Postdoctoral Science Foundation (20110490255), and the Beijing Postdoctoral Research Foundation (2011ZZ-18).

5. REFERENCES

[1] F. Liu, C. T. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):28–40, 2004.

[2] Y. Ma, Y. Zeng, X. Ren, and N. Zhong. User interest modeling based on multi-source personal information fusion and semantic reasoning. In *Proceedings of the 2011 International Conference on Active Media Technology (AMT 2011)*, pages 195–205, 2011.

[3] K. Sugiyama and M.-Y. Kan. Scholarly paper recommendation via user's recent research interests. In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL 2010)*, pages 29–38, 2010.

[4] Z. Wen and C.-Y. Lin. How accurately can one's interests be inferred from friends. In *Proceedings of the 19th international conference on World Wide Web (WWW 2010)*, pages 1203–1204, 2010.

[5] Y. Zeng. Interest-driven unification of retrieval and reasoning: Its theory and applications. Postdoctoral research report, Beijing University of Technology, 2012.

[6] Y. Zeng, N. Zhong, Y. Wang, Y. Qin, Z. Huang, H. Zhou, Y. Yao, and F. van Harmelen. User-centric query refinement and processing using granularity based strategies. *Knowledge and Information Systems*, 27(3), 2011.

[7] Y. Zeng, E. Zhou, Y. Wang, X. Ren, Y. Qin, Z. Huang, and N. Zhong. Research interests : Their dynamics, structures and applications in unifying search and reasoning. *Journal of Intelligent Information Systems*, 37(1):65–88, 2011.