DYNAMIC PROFILE REPRESENTATION AND MATCHING IN DISTRIBUTED SCIENTIFIC NETWORKS

Pham Tran Vu
University of Technology, VNU-HCM

(Manuscript Received on Decmber 07th, 2010, Manuscript Revised April 21st, 2011)

ABSTRACT: *Finding people having similar interests in online community is an interesting but challenging problem. Especially, in distributed multidisciplinary research network, locating scientists who share common interests to collaborate and solve large scientific problems is becoming more important. This paper introduces a method for extracting and modeling scientists' interest profiles from their day-to-day interactions and a method for semantically matching interest profiles based on latent semantic analysis. These methods exclude the necessity of having ontology for semantic matching of profiles, while still maintain the ability to reason about the semantic meaning of words.*

Keywords*: Scientific network, semantic matching, profile representation.*

1. INTRODUCTION

Finding people having similar research interests to initiate scientific collaboration is becoming important in distributed scientific communities. It is especially more essential in communities, where interdisciplinary research collaboration play an integral part. This is also a well-known research problem (commonly known as expert finding or profile matching) in information retrieval and has been studied by many previous researches.

The problem of finding people having similar interests is composed of two sub-problems: (i) how to profile interests of a person and, (ii) how to calculate the similarity between interest profiles.

In traditional online applications, e.g. news, users can explicitly specify their interests statically in their user profiles during registration. This approach is simple, but suffers from a couple of limitations. Firstly, the users may not realize exactly their interests. Secondly, a user's interests may change overtime. As the result, the initial profile registration will no longer be valid, unless it is updated regularly.

In recently years, different ways of building user profiles dynamically has been introduced. Instead of explicitly specified, profiles are implicitly extracted from different sources of information such as Wikipedia [1-3], citation analysis [4] and expert's documents [5]. User profiles extracted using these methods may well reflect the users' expertise over long period of time. However, without the time dimension, it cannot be used to conclude what the current interest of a user is, as interest and expertise are not always the same.

Profiles, once having been generated, can be matched using different methods. The most common method used in information retrieval for content matching is cosine similarity. In the methods, profiles are represented as a set of (weighted) keywords. The similarity is calculated by the cosine of the two vectors represented by two respected keyword sets. This method is simple to implement, but it only deals with syntactic matching of words. It is not sufficient in cases where comparing semantic meaning of words is required. To address the needs for semantic matching of user profiles, ontology can be used to explicitly describe relationships between words [6]. The relationships between words or terms can also be calculated implicitly, using latent semantic analysis [7, 8].

The emergence of social Web and Web 2.0 in recent years has created more opportunities for scientists to share resources and collaborate. In a social Web environment, scientists can share scientific resources, most popularly publications, with their peers. They can review and make comments on resources shared by others. We believe that the activities that a scientist performs on a social Web environment (e.g. sharing, reading, commenting and tagging papers) reflect his/her current interests. On this basis, we have developed a method for building scientific interest profiles implicitly. In this paper, we also introduce a our method of matching interest profiles by combing the tradition cosine similarity and latent semantic analysis techniques as suggested in [8]. We use of latent semantic analysis for semantic matching to avoid the necessity of an explicit ontology. In a multilingual and multidisciplinary research environment, it is almost impossible to have ontology that covers the knowledge of the whole environment.

2. APPLICATION CONTEXT

Motivated by the current success of social Web such as Facebook, Youtube, and Linkedin, we are building a social Web based virtual research environment, which allows scientists to:

- Share research ideas, documents, tools and data
- Locate expertise and set up network for solving scientific problems
- Set up and manage group activities
- Get access to high end computational resources.

The major goal is to have an environment in which scientists from different research disciplines can participate, share knowledge and collaborate to solve large scientific problems. Find people with similar interests in one of the core function of this environment.

3. PROFILE REPRESENTATION

A scientific profile consists of many different types of information, including the scientist's education background, work information, achievements and awards. These types of information of course can also be used to infer the scientist's research interests. However, they are static and may not well reflect the scientist's interests over time. This paper is focused on the dynamic information that can be used to extract the scientist's current interests.

3.1. Interaction Model

The interaction model used to extract scientists' interests is described in Figure 1. This is a three way relation between user, resource and tag. The interaction between a user and a resource can be in form of uploading, accessing, modifying, commenting or tagging. A user can give tags to a resource. In this interaction model, the user's interests are inferred in using the following assumptions:

- A user interacts with a resource implying that the user has an interest in that resource. The frequency of interaction implies the intensity of the interest.
- A user gives a tag to a resource implying that the user is interested in the content described by the meaning of the tag. The association frequency between a user and a tag implies the intensity of the interest.
- A tag is assigned to a resource implying that the resource's content can be described by the meaning of the tag. The frequency of the assignment implies the strength of the association.

All the interactions are time-stamped. When calculating a user's interests at a particular point of time, only interactions happened within a time window covering that point are used. In the following discussion, assuming that all the interacting happened within a single time window, the time dimension is not explicitly mentioned.

From the above interaction model, a user's interest profile can be modeled by a bag of weighted tags and a bag of weighted resources. Weights of tags and resources are normalized frequencies values of tags and resources respectively, using the following formula:

$$w_i = \frac{f_i}{\sum_i f_i} \qquad (1)$$

Where, f_i is the frequency of association between the user and a tag (or resource).

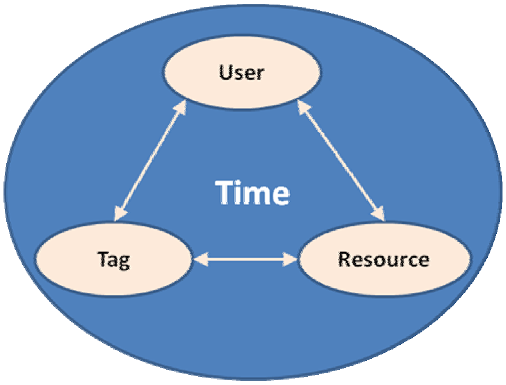

Figure 1. The interaction model – three way relation between user, resource and tag

Let T be the bag of weighted tags and R be the set of weighted resources, then T and R are sets of binary tuples:

$$T = \{(w_0^t, t_0), \ldots, (w_n^t, t_n)\} \qquad (2)$$

$$R = \{(w_0^r, r_0), \ldots, (w_m^r, r_m)\} \qquad (3)$$

Where, w_i^t and w_j^r are weights of tags and resources respectively. If u is the interest profile of a user, then:

$$u = \alpha T \cup \beta R \qquad (4)$$

Where, α and β are the relative contributions of the bag of tags and the bag of resources to the total user interest profile.

3.2. Resource Model

Tags can be directly used as terms in calculation. However, resources are complex objects. They need to be further decomposed.

Resources can be documents, research data sets, or scientific publications. A resource is often described by a title and a short description. For example, a research paper is often associated with a title, an abstract, and a set of keywords. Through the interactions within the environment, a resource may also be tagged. Terms are extracted from descriptions associated with a resource. The result of extraction and associated tags form a bag of weighted terms that describe the resource. The weights of terms are calculated from term frequencies using equation (1). Therefore, a resource can also be modeled as a set of binary tuples:

$$r = \{(w_0, t_0), ..., (w_k, t_k)\} \qquad (5)$$

Combining (2), (3), (4) and (5), the user interest profile can be generally represented as a set of binary tuples:

$$u = \{(W_0, t_0), ..., (W_p, t_p)\} \qquad (6)$$

Where, W_i is the aggregated weight of term t_i (terms and tags are treated the same way in this equation and referred to as terms generally in later discussions).

4. PROFILE MATCHING

Using user profile representation as in equation (6), the cosine similarity can be applied to calculate the similarity of any two user profiles. However, cosine calculation is limited to syntactic matching of terms. Semantic similarity is ignored in cosine calculation. This limitation can be overcome by combining semantic matching technique with cosine similarity. The terms that are in the intersection between the two profiles are used in cosine similarity calculation. Semantic matching technique is used for other terms. The final result is the aggregation of the two calculations [6].

4.1 Semantic Analysis

The semantic of terms can be defined explicitly using ontology. It gives dictionary-like definitions and relationships between terms. However, in a multidisciplinary research environment, it is difficult to have a common ontology that cover all domains. In this work, we apply latent semantic analysis technique [7] to extract the semantic relationships between terms. Our assumption is that if the two terms (or tags) happen to be in the same resource, they somehow relate to each other in meaning.

A term-resource matrix of size $|t| \times |r|$ is constructed to hold information about the weighted occurrences of terms in resources.

The value at row i and column j represents the normalized weight of term i in resource j as in equation (5). Each row of the matrix is a vector showing weighted occurrences of a term in all resources. The normalized dot product of any two rows is the occurrence correlation of any two terms. It is the implicit semantic relationship we use for semantic matching.

$$Sim(t_i,t_k)=\frac{\sum_j w_{i,j}w_{k,j}}{\sqrt{\sum_j (w_{i,j})^2}\times\sqrt{\sum_j (w_{k,j})^2}} \qquad (7)$$

4.2 Semantic Matching

Given two user profiles u and v represented by two sets of binary tuples. The similarity of u and v is calculated as:

$$Sim(u,v)=\frac{N_{\cos}}{N}Sim_{\cos}(u,v)+\frac{N_{sem}}{N}Sim_{sem}(u,v)$$

$Sim_{\cos}(u,v)$ is the cosine similarity calculated using the set of terms that in the intersction terms in u and terms in v. $Sim_{sem}(u,v)$ is the semantic similarity calculation for the non-overlapping part. N, $N_{\cos}$ and N_{sem} are the total number of terms of the two profiles, the number of terms involved in cosine and semantic calculations, respectively.

Let u' and v' be the overlapping portions of u and v, respectively, then:

$$u'=\{(w_0^{u'},t_0),\ldots,(w_k^{u'},t_k)\}$$

$$v'=\{(w_0^{v'},t_0),\ldots,(w_k^{v'},t_k)\}$$

Where, $w_i^{u'}$ and $w_i^{v'}$ are the weights of term t_i in u' and v' respectively.

$$Sim_{\cos}=\frac{\sum_i w_i^{u'}w_i^{v'}}{\sqrt{\sum_i (w_i^{u'})^2}\times\sqrt{\sum_i (w_i^{v'})^2}} \qquad (8)$$

Calculation of $Sim_{sem}(u,v)$ is more complicated. Let u'' and v'' be non-overlapping portions of u and v, respectively. For each term in u'', its average similarity with all terms in v'' is calculated using equation (7). The sum of these average values is divided by the number of terms in u'' to get the semantic similarity, as in the following equation:

$$Sim_{sem}(u,v)=\frac{\sum_i\sum_j Sim(t_i^{u''},t_j^{v''})}{|u''|\times|v''|} \qquad (9)$$

5. CONCLUSION

This paper has introduced a method for dynamic extraction, building and representation of user interest profiles, and a method for semantic matching of user profiles. The key advantages of these methods are:

- The users do not need to explicitly specify and regularly update their interest profiles. The system will automatically learn and update them through time.
- Building ontology for across domain collaboration is a challenging problem. It is extremely hard in multilingual environments. The profile matching method introduced is able to deal with semantic meaning of terms, but do

not need to use any ontology. This helps to reduce the complexity of developing and maintaining ontology.

In addition to building and matching user profiles, the methods presented can also be applied to other application areas of information retrieval such as content filtering and recommendation.

BIỂU DIỄN VÀ SO SÁNH ĐỘNG HỒ SƠ CÁ NHÂN TRONG CÁC MẠNG KHOA HỌC

Phạm Trần Vũ

Trường Đại học Bách Khoa, ĐHQG-HCM

TÓM TẮT: *Tìm kiếm những người có cùng sở thích trong các cộng đồng mạng trực tuyến là một bài toán khó và hấp dẫn. Đặc biệt, đối với các cộng đồng nghiên cứu đa ngành bị cách trở về mặt địa lý, việc tìm ra những người có cùng mối quan tâm để giải quyết các tài toán khoa học lớn ngày càng quan trọng. Bài báo này giới thiệu một phương pháp tổng hợp hồ sơ mối quan tâm của các nhà khoa học thông qua quá trình tương tác của họ trên cộng đồng, và phương pháp so trùng các hồ sơ dựa trên các phân tích về mặt ngữ nghĩa. Các phương pháp này không cần sử dụng ontology, nhưng vẫn có khả năng thực hiện các so sánh liên quan đến ngữ nghĩa, dựa vào các phương pháp thống kê.*

Từ khóa*: Mạng khoa học, so sánh ngữ nghĩa, biểu diễn hồ sơ cá nhân.*

TÀI LIỆU THAM KHẢO

[1]. G. Demartini, "Finding Experts Using Wikipedia," presented at FEWS, 2007.

[2]. E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," presented at IJCAI, 2007.

[3]. S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using Wikipedia," presented at SIGIR, 2007.

[4]. T. Bogers, K. Kox, and A. van den Bosch, "Using Citation Analysis for Finding Experts in Workgroups," presented at DIR, 2008.

[5]. H. Jung, M. Lee, I. S. Kang, S. Lee, and W. K. Sung, "Finding topic-centric identified experts based on full text analysis," presented at FEWS, 2007.

[6]. Rajesh Thiagarajan, Geetha Manjunath, and M. Stumptner, "Finding Experts By Semantic Matching of User Profiles," presented at Personal Identification and Collaborations: Knowledge Mediation and Extraction, 2008.

[7]. T. K. Landauer, P. W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," Discourse Processes, vol. 25, pp. 259-284, 1998.

[8]. B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme, "Evaluating Similarity Measures for Emergent Semantics of Social Tagging," presented at WWW, Madrid, 2009.