An Algorithmic Framework for Collaborative Interest Group Construction

Akshi Kumar, Abha Jain

Dept. of Computer Engineering, Delhi Technological University, Delhi, India
{akshikumar@dce.ac.in, abhajain87@yahoo.com}

Abstract. As organizations, both business and research-development continue to search better ways to exploit knowledge capital accumulated on the diversified Web; it fosters the need of collaboration among people with similar interest & expertise. In this paper we focus on the problem of discovering people who have particular interests or expertise. The standard approach is to build interest group lists from explicitly registered data. However, doing so assumes one knows what lists should be built, and who ought to be included in each list. We present an alternative approach, which can support a finer grained and dynamically adaptive notion of shared interests. Our approach deduces shared interest relationships between people based on interest similarity calculated by the means of entries written on their blog. Using this approach, a user could search for people by requesting a list of people whose interests are similar to several people known to have the interest in question.

Keywords: Expert finding, Interest Similarity, Blogs, Interest Groups.

1. Introduction

Ongoing increases in wide-area network connectivity promise vastly augmented opportunities for collaboration and resource sharing. A fundamental problem that confronts users of such networks is how to discover the existence of resources of interest, such as files, retail products, network services, or people. In this paper we focus on the problem of discovering people who have particular interests or expertise. We propose an Interest Group construction algorithm based on interest similarity, which can cluster researchers with similar interests into the same group and facilitate collaborative work.

Until very recently, finding expertise required a mix of individual, social and collaborative practices, a haphazard process at best. Mostly, it involved contacting individuals one trusts and asking them for referrals, while hoping that one's judgment about those individuals is justified and that their answers are thoughtful. The traditional way of providing expert assistance relied on the building a directory through manual entry of expertise data or explicitly registered data, such as X.500 directory service standard [1] or Microsoft's SPUD [2] etc. While this approach provides good support for locating particular users (the "white pages" problem), it does not easily support finding users who have particular interests or expertise (the

"yellow pages" problem). Moreover manual collecting method requires intensive and expensive labor, and may quickly outdate due to the continuous change in people's specific expertise and skills. Recently, more attention is paid on automated systems that enhance the visibility and traceability of experts [3] [4] [5] [6] [7] [8]. These systems aim at mitigating the above shortcomings by trying to automatically discover up-to-date expertise from implicit/secondary sources instead of relying on experts and/or other human sources. The information source for these systems includes electronic mail [9], discussion groups [6] [10], personal web pages [4] [5], Web browsing pattern [3] [8], various documents/reports related to particular users [11] [12], etc.

With the advent of Web, a new type of collaborative edifying & learning method has come into being. The traditional method to learning took the teacher as the core of the system and inevitably created many "*Forlorn & Lonely"* learners. So, it is necessary to promote the cooperative activities among the members of the teaching & research communities. To solve the problem, we need to organize individuals with the same interests into the same group, so as to help them carry out cooperative research work & learning. For the reasons above, based on the idea of interest similarity, we have put forward a constructing algorithm of Interest Group.

The demand for knowledge management, including methodologies for enhancing the intellectual faculties of an organization or a community, is increasing. An important factor in knowledge management is finding a person who has a high level of expertise in a required area. In general, an expert is someone who possesses a high level of knowledge in a particular domain. This implies that experts are reliable sources of relevant resources and information. However, the conventional way of doing this relies on connections between individuals. It goes without saying that a more systematic way is required. Recently, many people have started to write their documents in electronic forms such as word processor files, e-mail messages or blogs. For engineers and researchers, this has meant a lot of their expertise written in such documents. Therefore, analyzing these documents would make it possible to estimate their expertise.

From a humble beginning as 'What's New' pages, blogs have arisen to become arguably the most popular online personal publishing platform on the internet. Many users search and read the blog sites to get grass-roots opinions, new-product evaluations, and so on. As a consequence of this trend, there are many web services that analyze blog documents and show recent topics [13]. In this paper, we put forward an approach, which deduces shared interest relationships between researchers based on the entries written on his/her blog and discuss how to extract, build and match individual researcher's interest from their blog document entries & finally detect their level of expertise in that research area. Further we organize the researchers with the same interests into the interest groups, so as to help them carry out collaborative work.

2. Determination of the Interest Similarity Relations

The difficulty and key point of constructing an Interest Group is to determine and calculate the similarity relations. This includes two steps, one is getting the dominant indication (just the Interest Vector) from the interests' recessive indication; another is calculating the Interest Similarity.

2.1 The Interest Vector

Each researcher writes blog entries according to his or her interest. Thus, it can be supposed that terms related to the researcher's interests are present in many entries in his or her blog site. The interest vector of the researcher, Vi, is represented as a bag-of-words with frequently used words being assigned high weights. The interest vector is calculated by the equation described below:

$$V_i = (s_{i1}, s_{i2}, s_{i3}, ...) \quad \textbf{(1)}$$

$$s_{ik} = ef_i(w_k) \times \log\left(\frac{N_u}{uf(w_k)}\right) \quad \textbf{(2)}$$

where s_{ik} means the strength of interest in word w_k; $ef_i(w_k)$ means the number of entries containing w_k in researchers i's site; $uf(w_k)$ means the number of researchers who use w_k; and N_u means the number of researchers. This equation corresponds to the traditional *tf-idf* weighting approach. The entry frequency, $ef_i(w_k)$, corresponds to *tf*, and inverse user frequency, $N_u/uf(w_k)$, corresponds to *idf*. Thus, a word repeatedly used in a small number of blog sites has high weight value.

2.2 Similarity scores between Researchers

A similarity score represents how similar the interests of a pair of researchers are. If researcher *i* and *j* have similar interests, their interest vectors should be similar. Thus, we calculate the similarity score between them, Rij, using the cosine similarity of Vi and Vj as described below.

$$R_{ij} = \frac{V_i \times V_j}{|V_i||V_j|} \quad \textbf{(3)}$$

All elements of Vi and Vj are positive and thus the range of Rij is 0 to 1.

3. Assessing Expertise: Why it matters?

We seek *guidance* from people who are familiar with the choices we face, who have been helpful in the past, whose perspectives we value, or who are recognized experts. In general, an expert is someone who possesses a high level of knowledge in a particular domain. This implies that experts are reliable sources of relevant resources and information. Following expert users provides more benefits:

- ❖ Should know the best resources with respect to a given topic.
- ❖ Should be quick in discovering and identifying new resources

An open problem thus arises to how can level of expertise be assessed objectively? We propose the solution for this by calculating every researcher's level of expertise [e] (that is the number of the researchers who have high interest similarity with a specific researcher).
Suppose there are m researchers, the researcher i's level of expertise will be calculated by the following formula:

$$e_i = \frac{1}{m}\sum_{j=1}^{m} ac_{ij} \qquad \textbf{(4)}$$

In this Formula,

$$ac_{ij} = \begin{cases} 1 & \text{if } R_{ij} \geq T_1 \\ 0 & \text{otherwise} \end{cases}$$

T_1 is a pre-determined Threshold Value

4 Algorithm for Construction of Collaborative Research Interest Group

A Collaborative Research Interest Group should be a group consists of researchers in the similar area or with related interests. So, when constructing a group, try to arrange the researchers with great interest similarity into the same group. With this theory, we put forward the steps for constructing the Interest Group. The proposed method has 4 steps. Firstly we extract the researchers' Interest Vector from their blog documents; and then, with the Interest Vectors, we calculate the Interest Similarity between two researchers. Next, we compute the Level of expertise to find the experts in area and lastly, with these data, we construct an Interest group in a certain way.

Step 1: Use formula (3) to calculate the interest similarity between two researchers.

Step 2: Calculate every researcher's level of expertise [e], i.e., the number of the researchers who have high interest similarity with a specific researcher, using (4).

Step 3: Select the researcher with the highest Level of Expertise, and take him/her as the center of the group to be constructed. Pre-determine a threshold value, T_1, those researchers whose interest similarities with the centered researcher are higher than the threshold value can access into the group.

Step 4: As for the rest of the researchers, recalculate according to the step 1 to step 3, until the researchers' highest level of expertise is less than the threshold value T_1, then, stop calculating.

Some additional points to be explained:

- Because the interests of researchers are distributed randomly, some groups may have many members. We are not setting any restrictions on the number count and letting a pre-determined threshold value control the number of members in the interest group.
- In the process of constructing interest groups, we should consider that, the interests of the members in the group are in dynamic changing. So the conditions of the group are also dynamically changing. When constructing groups, we can save the individuals interest property value as the groups' core values. Once an researcher's interest changed, calculate his/her instant interest value's similarity with relative core value. If the similarity value is less than the pre-determined threshold value, let the researcher withdraw from the community and recalculate which communities should he/she go.
- When constructing communities, some researchers have many different kind of interests, there may be one researcher belongs to several communities at the same time. This means he/she can take part in the activities in several communities.

5. Case Study

To clearly illustrate the effectiveness of the proposed algorithm for Construction of Collaborative Research Interest Group, a case study is presented to describe a typical scenario, where

- There are 5 researchers viz. i, j, k, n & m. Therefore, $N_u = 5$
- There are 5 entries in each of the researcher's blog site.

The following table 1 shows the blog entries of each of the researcher i, j, k, n & m.

Table 1. Sample blog entries of 5 researchers

Researcher / Entry	i	j	k	n	m
1	w_1, w_{16}, w_3, w_2, w_{17}, w_9	w_{14},w_8, w_6, w_7, w_{17}	w_{11}, w_7, w_2, w_9, w_{19}	w_{13}, w_{13}, w_{10}, w_{14}	w_{10},w_{15}, w_2
2	w_4, w_2,w_3, w_{14}, w_{11}, w_{18}	w_1, w_{16}, w_{11}, w_7, w_{18}, w_{17},w_6	w_{14},w_{10},w_4, w_9,w_{19}	w_{11},w_{13}, w_6, w_5	w_{14},w_{16}, w_9, w_8
3	w_1,w_2,w_6, w_{13}	w_7, w_3, w_{18},w_8, w_{17}	w_9, w_{19}, w_{11},w_{10}, w_{17}	w_{13}, w_{14}, w_{18}, w_{12}	w_{15},w_{19}, w_1, w_{16}
4	w_1,w_2, w_4, w_8,w_{15}, w_{10}	w_6, w_6,w_7, w_{17}	w_{12}, w_9, w_{19},w_{16}	w_{17}, w_{13}, w_2	w_{11},w_{17}, w_6, w_{15}
5	w_1,w_2,w_5, w_3, w_{19}	w_7,w_{18},w_{15}, w_2,w_{18} ,w_6, w_{17}, w_1	w_{19}, w_9,w_1, w_{17}, w_{10}, w_{10}	w_{18},w_7, w_{13},w_{13}	w_3, w_{13}

5.1 Interest Vector Calculations

We have the interest vector corresponding to each of the researcher i, j, k, n & m represented as V_i, V_j, V_k, V_n, V_m. The calculation for these vectors using equation 2 is shown below:

For Researcher i: The Interest Vector is: $V_i = (S_{i1}, S_{i2}, S_{i3}, S_{i4}, S_{i5})$ where ;

S_{i1}=ef (w_1) x log [5 / uf (w_1)]
S_{i2}=ef (w_2) x log [5 / uf (w_2)]
S_{i3}=ef (w_3) x log [5 / uf (w_3)]
S_{i4}=ef (w_4) x log [5 / uf (w_4)]
S_{i5}=ef (w_5) x log [5 / uf (w_5)]

Now, from table 1, we find the values for ef's and uf's for the corresponding words:

ef (w_1)=4 ; uf (w_1)=3 => S_{i1} = 4 * log (5/3) = 0.8874

ef (w_2)=5 ; uf (w_2)=4 => S_{i2} = 5 *log (5/4) = 0.4846

ef (w_3)=3 ; uf (w_3)=2 => S_{i3} = 3*log (5/2) = 1.1938

ef (w_4)=2 ; uf (w_4)=1 => S_{i4} = 2*log (5/1) = 1.3979

ef (w_5)=1 ; uf (w_5)=1 => S_{i5} = 1*log (5/1) = 0.6989

Thus, V_i= (0.8874, 0.4846, 1.1938, 1.3979, 0.6989)

For Researcher j: The Interest Vector is: $V_j = (S_{j6}, S_{j7}, S_{j8}, S_{j17}, S_{j18})$ where;

S_{j6}=ef (w_6) x log [5 / uf (w_6)]
S_{j7}=ef (w_7) x log [5 / uf (w_7)]
S_{j8}=ef (w_8) x log [5 / uf (w_8)]
Sj_{17}=ef (w_{17}) x log [5 / uf (w_{17})]
Sj_{18}=ef (w_{18}) x log [5 / uf (w_{18})]

Now, from table 1, we find the values for ef's and uf's for the corresponding words

ef (w_6)=4; uf (w_6)=3 => S_6 = 4*log (5/ 3) = 0.8874

ef (w_7)=5; uf (w_7)=2 => S_7 = 5*log (5/2) =1.9897

ef (w_8)=2; uf (w_8)=2 => S_8 = 2*log (5/2) = 0.7959

ef (w_{17})=5; uf (w_{17})=4 => S_{17} = 5*log (5/4) = 0.4845

ef (w_{18})=3; uf (w_{18})=3 => S_{18} = 3*log (5/3) = 0.6655

Thus, Vj = (0.8874, 1.9897, 0.7959, 0.4845, 0.6655)

For Researcher k: The interest vector is: $V_k = (S_{k9}, S_{k10}, S_{k11}, S_{k12}, S_{k19})$ where;

S_{k9}=ef (w_9) x log [5 / uf (w_9)]
S_{k10}=ef (w_{10}) x log [5 / uf (w_{10})]
S_{k11}=ef (w_{11}) x log [5 / uf (w_{11})]
S_{k12}=ef (w_{12}) x log [5 / uf (w_{12})]
S_{k19}=ef (w_{19}) x log [5 / uf (w_{19})]

Now, from table 1, we find the values for ef's and uf's for the corresponding words
ef (w_9)=5; uf (w_9)=2 => S_9 = 5*log (5/2) = 1.9897

ef (w_{10})=3; uf (w_{10})= 3 => S_{10} = 3*log (5/3) = 0.6655

ef (w_{11})=2; uf (w_{11})=4 => S_{11} = 2*log (5/4) = 0.1938

ef (w_{12})=1; uf (w_{12})=1 => S_{12} = 1*log (5/1) = 0.6988

ef (w_{19})=5 uf (w_{19})=2 => S_{19} = 5*log (5/2) = 1.9897

Thus, V_k = (1.9897, 0.6655, 0.1938, 0.6988, 1.9897)

For Researcher n: The Interest Vector is: V_n = (S_{n13}, S_{n14}, S_{n20}, S_{n21}, S_{n22}) where;

S_{n13}=ef (w_{13}) x log [5 / uf (w_{13})]
S_{n14}=ef (w_{14}) x log [5 / uf (w_{14})]
S_{n20}=ef (w_{20}) x log [5 / uf (w_{20})]
S_{n21}=ef (w_{21}) x log [5 / uf (w_{21})]
S_{n22}=ef (w_{22}) x log [5 / uf (w_{22})]

Now, from table 1, we find the values for ef's and uf's for the corresponding words

ef (w_{13})=5; uf (w_{13})=2 => S_{13} = 5*log (5/2) = 1.9897

ef (w_{14})=2; uf (w_{14})=4 => S_{14} = 2*log (5/4) = 0.1938

ef (w_{20})=4; uf (w_{20})=3 => S_{20} = 4*log (5/3) = 0.8874

ef (w_{21})=3; uf (w_{21})=4 => S_{21} = 3*log (5/4) = 0.2907

ef (w_{22})=4; uf (w_{22})=2 => S_{22} = 4*log (5/2) = 0.8874

Thus, V_n = (1.9897, 0.1938, 0.8874, 0.2907, 0.8874)

For Researcher m: The Interest Vector is: V_m = (S_{m15}, S_{m16}, S_{m23}, S_{m24}, S_{m25}) where;

S_{m13}=ef (w_{15}) x log [5 / uf (w_{15})]
S_{m14}=ef (w_{16}) x log [5 / uf (w_{16})]
S_{m20}=ef(w_{23}) x log [5 / uf (w_{23})]
S_{m24}=ef(w_{24}) x log [5 / uf (w_{24})]
S_{m25}=ef(w_{25}) x log [5 / uf (w_{25})]

Now, from table 1, we find the values for ef's and uf's for the corresponding words

ef (w_{15})=3; uf (w_{15})=2 => S_{15} = 3*log (5/2) = 1.1938

ef (w_{16})=2; uf (w_{16})=3 => S_{16} = 2*log (5/3) = 0.4436

ef (w_{23})=4; uf (w_{23})=4 => S_{23} = 4*log (5/4) = 0.3876

ef (w_{24})=5; uf (w_{24})=4 => S_{24} = 5*log (5/4) = 0.4845

ef (w_{25})=2; uf (w_{25})=4 => S_{25} = 2*log (5/4) = 0.1938

Thus, V_m = (1.1938, 0.4436, 0.3876, 0.4845, 0.1938)

5.2 Similarity Score Calculation

Using the formula defined in equation 3, we calculate the values of Similarity Score between each of the 2 researchers:

R_{ij} = 0.7063; R_{ik} = 0.7110; R_{in} = 0.7502; R_{im} = 0.8064; R_{jk} = 0.6688; R_{jn} = 0.6132
R_{jm} = 0.7424; R_{kn} = 0.8786; R_{km} = 0.8140; R_{nm} = 0.9169

As all the elements of both the vectors taken at a time to calculate the similarity score are positive, thus the range of similarity score is between 0 to 1.

This indicates that:

- The value of 1 means that the 2 researchers have exactly similar interests and;
- The value of 0 means that the 2 researchers do not have any similar interests at all.

Therefore, we can conclude that:

- The researchers n & m have almost similar interests (as R_{nm}= 0.9169, approx 1)
- The researchers k & n have similar interests to a very great extent (as R_{kn} = 0.8786)
- The researchers “k & m” and “i & m” have quite a lot similar interests (as R_{km} = 0.8140 and R_{im} = 0.8064)
- The researchers “j & k” and “j & n” have quite less similar interests (as R_{jk} = 0.6688 and R_{jn} = 0.6132)

6. Conclusion

This paper expounds an entirely different approach to solve the problem of discovering people who have particular interests or expertise. We have put forward a constructing algorithm of Interest Group by uncovering shared interest relationships between people, based on their blog document entries, to let them arrange into groups effectively, to let them share the resources, carry out cooperative work. The practice result proves that this algorithm has the characteristics of highly effective group arranging and is easy to be extendable.

References

1. CCITT/ISO. The Directory, Part 1: Overview of Concepts, Models and Services. CCITT/ISO, Gloucester, England, CCITT Draft Recommendation X.500/ISO DIS 9594-1(1988).

2. Davenport T. H. and Prusak L., Working Knowledge: How Organizations Manage What They Know, Boston, MA: Harvard Business School Press (1998).

3. Cohen A. L., Maglio P. P., Barrett R.: The Expertise Browser: How to Leverage Distributed Organizational Knowledge, presented at Workshop on Collaborative Information Seeking at CSCW'98, Seattle, WA (1998).

4. Kautz H., Selman B. and Shah M.: The Hidden Web, The AI Magazine, vol. 18, no. 2, pp. 27 – 36 (1997).

5. Kautz H., Selman B. and Milewski A.: Agent Amplified Communication, in Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), Portland, OR, pp. 3 – 9 (1996).

6. Krulwich B. and Burkey C. : ContactFinder: Extracting Indications of Expertise and Answering Questions with Referrals, in Working Notes of the 1995 Fall Symposium on Intelligent Knowledge Navigation and Retrieval, Cambridge, MA. Technical Report FS-95-03, The AAAI Press, pp. 85 – 91 (1995).

7. Mattox D., Maybury M. and Morey D.: Enterprise Expert and Knowledge Discovery, in Proceedings of the 8th International Conference on Human-Computer Interaction (HCI International'99), Munich, Germany, pp. 303-307 (1999).

8. Pikarakis A., et al, MEMOIR: Software Agents for Finding Similar Users by Trails, in Proceedings of the Third International Conference on the Practical Applications of Intelligent Agents and multi-Agent Technology (PAAM-98), London, UK, pp. 453 – 466 (1998).

9. Schwartz M. F. and Wood D. M.: Discovering Shared Interests Using Graph Analysis, Communications of the ACM, vol. 36, no. 8, pp. 78 – 89 (1993).

10. Krulwich B. and Burkey C.: The ContactFinder Agent: Answering Bulletin Board Questions with Referrals, in Proceedings of the 1996 National Conference on Artificial Intelligence (AAAI-96), Portland, OR, vol. 1, pp. 10 –15(1996).

11. Steeter L. A. and Lochbaum K. E.: An Expert/Expert Locating System based on Automatic Representation of Semantic Structure, in Proceedings of the Fourth IEEE Conference on Artificial Intelligence Applications, Computer Society of the IEEE, San Diego, CA, pp. 345 – 349 (1988).

12. Steeter L. A. and Lochbaum K. E., Who Knows: A System Based on Automatic Representation of Semantic Structure, in RIAO'88, Cambridge, MA, pp. 380 – 388(1988).

13. BlogPulse, http://www.blogpulse.com/