References

1. Blair, D.C., Maron M.E.: 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28 (1985)
2. van Rijsbergen, C., Information Retrieval, (1979)
3. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. Communications of the ACM, 30(11):964-971, November (1987).
4. Merkl, D., Exploration of Text Collections with Hierarchical Feature Maps (1997)
5. Rauber, A., Dittenbach, M., and Merkl, D., Automatically Detecting and Organizing Documents into Topic Hierarchies: A Neural Network Based Approach to Bookshelf Creation and Arrangement (2000)
6. Kohonen, T., Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43:-69, 1982.
7. Krista, L., Honkela, T., Kaski, S., and Kohonen, T., WEBSOM - A Status Report (1996)
8. Honkela, T., Pulkki, V., and Kohonen, T. (1995). Contextual relations of words in Grimm tales analyzed by self-organizing map. In Fogelman-Soulié, F. and Gallinari, P., editors, Proceedings of the International Conference on Artificial Neural Networks, ICANN-95, volume 2, pages 3-7, Paris. EC2 et Cie.
9. Kohonen, T., Kasaki., S., Langus., K., Salojärvi, J., Paatero., V. and Saarela, A. Self Organization of a Massive Document Collection. IEEE Transactions on Neural Networks for Data Mining and Knowledge Descovery, Volume 11 (3), pp 574-585. (2000)
10. Blackmore, J., Miikkulainen, R.: Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map. In Proc Int'l Conf Neural Networks (ICANN'93), San Francisco, CA, 1993.
11. Fritzke, B.: Growing grid - a self-organizing network with constant neighborhood range and adaption strength. Neural Processing Letters, 2, No. 5:1 - 5, (1995)
12. Chen, H., Houston., A., Sewell, R., Scatz., B., Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques (1998)
13. Salton, G., Wong, A., and Yang, C., Vector space model for automatic indexing, Communications of the ACM 18, pp. 613--620, 1975.
14. Rauber, A., Merkl, D., Automatic Labeling of Self-Organizing Maps: Making a Treasure-Map Reveal its Secrets
15. Freeman, R., Yin, H., Allinson, N., Self-Organising Maps for Tree View Based Hierarchical Document Clustering, Proceedings of the International Joint Conference on Neural Networks (IJCNN'02), Honolulu, Hawaii, vol. 2, pp. 1906-1911, (2002)

Natural Language Processing for Expertise Modelling in E-mail Communication*

Sanghee Kim[1], Wendy Hall[1], and Andy Keane[2]

[1]Intelligence, Agents, Multimedia Group, Department of Electronics and Computer Science
University of Southampton, U.K.
{sk98r,wh}@ecs.soton.ac.uk
[2]Computational Engineering and Design Center, School of Engineering Science
University of Southampton, U.K.
ajk@soton.ac.uk

Abstract. One way to find information that may be required, is to approach a person who is believed to possess it or to identify a person who knows where to look for it. Technical support, which automatically compiles individual expertise and makes this accessible, may be centred on an expert finder system. A central component of such a system is a user profile, which describes user expertise level in discussed subjects. Previous works have made attempts to weight user expertise by using content-based methods, which associate the expertise level with the analysis of keyword usage, irrespective of any semantic meanings conveyed. This paper explores the idea of using a natural language processing technique to understand given information from both a structural and semantic perspective in building user profiles. With its improved interpretation capability compared to prior works, it aims to enhance the performance accuracy in ranking the order of names of experts, returned by a system against a help-seeking query. To demonstrate its efficiency, e-mail communication is chosen as an application domain, since its closeness to a spoken dialog, makes it possible to focus on the linguistic attributes of user information in the process of expertise modelling. Experimental results from a case study show a 23% higher performance on average over 77% of the queries tested with the approach presented here.

1 Introduction

A crucial task in the distributed environments that most organizations operate is to effectively manage the useful knowledge held by individuals. Not only does this supplement additional resource, but it also contributes timely and up-to-date procedural and factual knowledge to enterprises. In order to fully maximize individually held resources, it is necessary to encourage people to share such valuable data. As their expertise is accumulated through task achievement, it is also important

* This work was funded by the University Technology Partnership (UTP) for Design, which is a collaboration between Rolls-Royce, BAE Systems and the Universities of Cambridge, Sheffield and Southampton.

H. Yin et al. (Eds.): IDEAL 2002, LNCS 2412, pp. 161-166, 2002.

exploit it as it is created. Such an approach allows individuals to work as normal thout demanding changes in working environments [6].

An expert finder is a system designed to locate people who have 'sought-after owledge' to solve a specific problem. It answers with the names of potential helpers ainst knowledge seeking queries, in order to establish personal contacts which link vices to experts. The ultimate goal of such a system is to create environments here users are aware of each other, maximizing their current resources and actively changing up-to-date information. Although the expert finder systems cannot always enerate correct answers, bringing the relevant people together provides opportunities r them to become aware of each other, and to have further discussions, which may ncover hidden expertise.

In designing technical support to maximize the use of such personal expertise, two sues have to be addressed; 1) how to simulate the personal contacts found in real nvironments, and 2) how to capture personal expertise while allowing users to work s they normally do without demanding changes in working environments. The xploitation of e-mail communication, which can be enhanced as a communication-ased learning tool, where individual experiences are shared among communicators, s proposed as an enabling technology. E-mail communication has become a major neans of exchanging information and acquiring social or organisational relationships, mplying that it would be a good source of information about recent and useful cooperative activities among users. It is hypothesized that because of its popularity, nformation mined from e-mail communication can be considered as information rom expertise discovery sources [3; 6]. In addition, as it represents an every day activity, it requires no major changes to working environments, which makes it suitable as a test environment.

A decision about whether an individual is an expert for a given problem may be made by consulting user profiles. Drawn from information retrieval studies, the frequencies of keywords have been extensively used for extracting user information from exchanged e-mail messages. However, there are at least three reasons why such an approach is inadequate when applied to expertise modelling. First, counting keywords is not adequate for determining whether a given document is factual information or contains some level of author expertise. Secondly, without understanding the semantic meanings of keywords, it is possible to assume that different words represent the same concept and vice versa, which triggers the retrieval of non-relevant information. Finally, it is not easy to distinguish question-type texts from potential answer documents, which support retrieval of the relevant documents for the given query. In addition, the argument that user expertise is action-centred and is often distributed in the individual's action-experiences, is the motivation behind work that relies on linguistic-oriented user modelling [2]. With this approach, when we regard given messages as the realization of involved knowledge, user expertise can be verbalized as a direct indication of user views on discussed subjects, and the levels of expertise are distinguished by taking into account the degree of significance of the words employed in the messages.

In this paper, a new expertise model, EMNLP (Expertise Modelling using Natural Language Processing) that captures the different levels of expertise reflected in exchanged e-mail messages, and makes use of such expertise in facilitating a correct ranking of experts, is presented. It examines the application of NLP (Natural Language Processing) techniques and user modelling to the development of an expert finder system based on e-mail communication. The creation of an expert finder system that can be embedded in a user's working environments enabling a prompt utilisation is one of the two main themes of this paper, and improving its competency values by using NLP for the profiling of users is the second.

2 Related Work

KnowledgeMail from Tacit Corp is the system most related to EMNLP, in that it adds an automatic profiling ability to some of existing commercial e-mail systems, to support information sharing through executing queries about the profiles constructed [6]. User profiles are formulated as a list of weight-valued terms by using one statistical method. A survey focusing on the system's performance reveals that users tend to spend extra time cleaning up their profiles in order to reduce false hits, which erroneously recommend them as experts due to unresolved ambiguous terms [3]. In an effort to reduce such problems, the application of NLP to profiling users is suggested. As a consequence, EMNLP is expected to generate more meaningful terms in user profiles.

The system described by Vivacqua et al. [8] model a user's programming skill by reading Java source code files, and analysing what classes or methods are used and how often. This result is then compared to the overall usage for the remaining users, to determine the levels of expertise for specific methods. Its automatic profiling and mapping of five levels of expertise (i.e., expert-advanced-intermediate-beginner-novice) are similar to those of EMNLP. However, the expertise assignment function is rather too simplified in so far as it disregards various coding patterns that might reveal the different skills of experts and beginners.

3 Descriptions of EMNLP

A design objective of EMNLP is to improve the efficiency of the task search, which ranks peoples' names in decreasing order of expertise against a help-seeking query. Its contribution is to turn once simply archived e-mail messages into knowledge repositories by approaching them from a linguistic perspective, which regards the exchanged messages as the realization of verbal communication among users. Its supporting assumption is that user expertise is best extracted by focusing on the sentence where users' viewpoints are explicitly expressed. NLP is identified as an enabling technology that analyzes e-mail messages with two aims; 1) to classify sentences into syntactical structures (syntactic analysis), and 2) to extract users' expertise levels using the functional roles of given sentences (semantic interpretation). Figure 1 shows the procedure for using EMNLP, i.e. how to create user profiles from the collected messages. Contents are decomposed into a set of paragraphs and heuristics (e.g., locating a full stop) are applied in order to break down each paragraph into sentences.

Syntactical analysis identifies the syntactic roles of words in a sentence by using a corpus annotation [1]. Apple Pie Parser is used and it is a bottom-up probabilistic chart parser [5]. The syntactical analysis supports the location of a main verb in a sentence, by decomposing the sentence into a group of grammatically related phrases.

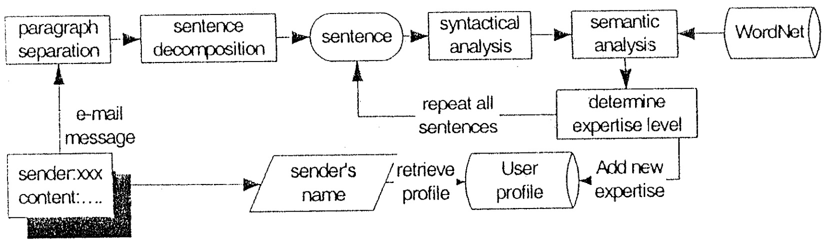


Fig. 1. The procedure for user profiling with the EMNLP

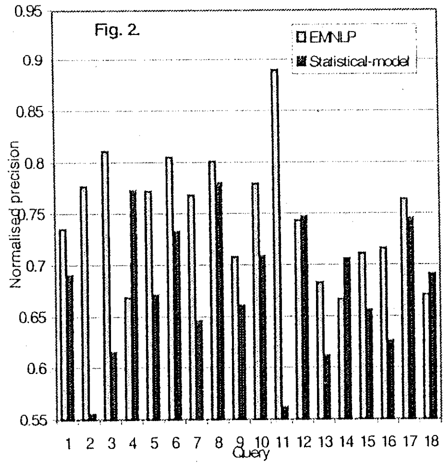
Given the structural information about each sentence, semantic analysis examines sentences with two criteria: 1) whether the employed verb verbalizes the speaker's attitudes, and 2) whether the sentence has a "first person" (e.g., "I", "In my opinion", or "We") subject. This analysis is based on Speech Act Theory (SAT), which proposes that communication involves the speaker's expression of an attitude (i.e. an illocutionary act) towards the contents of the communication [7]. It suggests that information can be delivered with different communication effects on recipients depending on different speaker's attitudes, which are expressed using an appropriate illocutionary act, which represents a particular function of communication. The performance of the speech act is described by a verb, which posits a core element as the central organizer of the sentence. In addition, the fact that working practices are reflected through task achievement implies that personal expertise can be regarded as action-oriented, emphasizing the important role of a "first person" subject in expertise modeling.

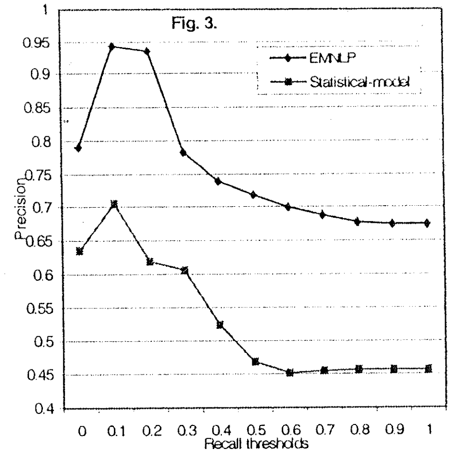
EMNLP extracts user expertise from the sentences, which have "first person" subjects, and determines expertise levels based on the identified main verbs. Whereas SAT reasons about how different illocutionary verbs convey the various intentions of speakers, NLP determines the intention by mapping the central verb in the sentence to the pre-defined illocutionary verb. The decision about the level of user expertise is made according to the defined hierarchies of the verbs, initially provided by SAT. SAT provides the categories of illocutionary verbs (i.e. assertive, commissive, directive, declarative, and expressive), each of which contains a set of exemplary verbs. EMNLP further extends the hierarchy in order to increase its coverage for practicability by using the WordNet Database [4]. EMNLP first examines all verbs occurring in the collected messages, and then filters out verbs, which have not been mapped onto the hierarchy. For each verb, it consults the WordNet database in order to assign a value through chaining its synonyms; for example, if the synonym of the given verb is classified into "assertive" value, and then this verb is also assigned into "assertive".

To clarify how two sentences, which may be assumed to contain similar keywords, are mapped onto different profiles, consider two example sentences: 1) "For the 5049 testing, phase analysis on those high frequency results is suggested", and 2) "For the 5049 testing, I know phase analysis on those high frequency results has to be added". The main verb values for both sentences (i.e., suggest and know) are equivalent to "working knowledge", which conveys a modest knowledge for a speaker. However, the difference is that when compared to the first, the second sentence clearly conveys the speaker's intention as it begins with "I know". As a consequence, it is regarded as demonstrating expertise while the first sentence is not. Information extracted from the first sentence is mapped onto a lower-level expertise.

4 Experimental Results

A case study has been developed to test two hypotheses; namely 1) that EMNLP produces comparable or higher accuracy in differentiating expertise from factual information compared to that of the frequency-based statistical model, 2) that differentiating expertise from factual information supports more effective query processing in locating the right experts. As a baseline, a frequency-based statistical model, which builds user profiles by weighting presented terms without considering their meanings or purposes, was used.





A total of 10 users, who work for the same department in a professional engineering design company, participated in the experiment and a period of three-to-four months duration was spent collecting e-mail messages. A total of 18 queries were created for a testing dataset, and a maximum number of 40 names of predicted experts, i.e. 20 names extracted using EMNLP and 20 names from the statistical model, were shown to a user, who was the group leader of the other users. As a manager, the user was able to evaluate the retrieved names according to the five pre-defined expertise levels: "Expert-Level Knowledge", "Strong Working Knowledge", "Working Knowledge", "Strong Working Interests", and "Working Interests".

Figure 2 summarizes the results measured by normalized precision. For 4 questions (i.e. 4,12,14,18), EMNLP produced lower performance rates than by using the statistical approach. However, for 14 queries, its ranking results were more accurate, and at the highest point, it outperformed the statistical method with a 33% higher precision value. The precision-recall curve, which demonstrates a 23% higher precision value for EMNLP, is shown in Figure 3. The differences of precision values at different recall thresholds are rather small with EMNLP, implying that its precision values are relatively higher than those of the statistical model.

A close examination of the queries used for testing reveals that the statistical model has a better capability in processing general-type queries that search for non-specific factual information, since 1) as we regard user expertise as action-oriented, knowledge is distinguished from such factual information, implying that it is difficult to value factual information as knowledge with EMNLP, and 2) EMNLP is limited to exploring various ways of determining the level of expertise in that it constrains user expertise to be expressed through the first person in a sentence.

5 Future Work

EMNLP was developed to improve the accuracy of ranking the order of expert names by use of the NLP technique to capture explicitly stated user expertise, which otherwise may be ignored. Its improved ranking order, compared to that of a statistical method, was mainly due to the use of an enriched expertise acquisition technique, which successfully distinguished experienced users from novices. We presume that EMNLP would be particularly useful when applied to large organizations where it is vital to improve retrieval performance since typical queries may be answered with a list of a few hundred potential expert names.

Special attention is given to gathering domain specific terminologies possibly collected from technical documents such as task manuals or memos. This is particularly useful for the semantic analysis, which identifies concepts and relationships within the NLP framework, since these terminologies are not retrievable from general-purpose dictionaries (e.g., the WordNet database).

References

1. Allen, J. (1987) Natural Language Understanding, Benjamin/Cummings Publishing
2. Choo, C. W., Detlor, B., Turnbull, D. (2000) WEB WORK Information Seeking and Knowledge Work on the World Wide Web, Kluwer Academic Publishers
3. Forbes (2001) Forbes, You've got expertise, http://www.forbes.com/global/2001/0205/088_print.html
4. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. (1993) Introduction to WordNet: An On-Line Lexical Database, University of Princeton, U.S.A.
5. Sekine, S., Grishman, R. (1995) A Corpus-based Probabilistic Grammar with only Two Non-Terminals, In Proceedings of the Fourth International Workshop on Parsing Technology, pp.216-223
6. Tacit (1999) White paper of KnowledgeMail, http://www.tacit.com/knowledgemail
7. Verschueren, J. (1980) On Speech Act Verbs, John Benjamins, Amsterdam
8. Vivacqua, A., Lieberman, H. (2000) Agents to Assist in Finding Help, In Proceedings of the CHI'2000, pp.65-72

A Branch and Bound Algorithm for Minimum Cost Network Flow Problem

Jun Han, Graham McMahon, and Stephen Sugden

School of Information Technology, Bond University
{jhan, gmcmahon, ssugden}@bond.edu.au

Abstract. In this paper we introduce a branch and bound algorithm for finding tree solution of minimum cost network flow problem. We consider different situations such as unit and non-unit traffic demand. The methods used to prune the searching tree in different situations are emphasized respectively but the complete searching process is not interpreted in detail due to limited space. Programming test results show the efficiency of these techniques.

1 Introduction

With the advent of the information age, there has been increased interest in the efficient design of communication networks. Considerable research interest has been shown in problems dealing with the economic synthesis of networks. The aim of this kind of design problem is to satisfy all traffic requirements at minimum cost between a set of nodes. The mathematical formulation of the problem can be as follows:

$$\text{Minimize} \sum_{p=1}^{n} \sum_{q>p}^{n} \sum_{r} h_r^{pq} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij} a_{ij,r}^{pq} \tag{1}$$

S.T.

$$\sum_r h_r^{pq} = F^{pq} \qquad \forall p,q \tag{2}$$

$$f_{ij} = \sum_{p=1}^{n} \sum_{q>p}^{n} \sum_{r} a_{ij,r}^{pq} h_r^{pq} \qquad \forall i,j \tag{3}$$

$$0 \le f_{ij} \le f_{ij}^{\max} \qquad \forall i,j \tag{4}$$

$$0 \le h_r^{pq} \qquad \forall p,q,r \tag{5}$$

where

n is the number of nodes,

h_r^{pq} is the amount of traffic on route r between O-D pair $p-q$,

H. Yin et al. (Eds.): IDEAL 2002, LNCS 2412, pp. 167-172, 2002.