1. INTRODUCTION

1.1. Background

Finding an expert in an organization is significant because many organizations are giving more emphasis to expertise of their knowledge (Dawit Y., 2000). Expert finding has to be able assist not only access to documented knowledge but the most important is knowledge held by individuals. The task of expert discovery is focused at identifying members of an organization with relevant expertise, skill, knowledge or experience for a given topic. Some systems can be assisted for organizations store and retrieve information about their experts who have the capability to perform specific tasks.

Discovering experts who have the appropriate skills and knowledge for a specify research field is a crucial tasks in academic activities. In an academic institute, there are needs to find specific information about academic staffs and researchers, for example, a potential postgraduate student who is looking for a suitable supervisor. The experts will publish some conference and journal based on their expertise. However, the experts’ information may change over the time. Thus, the skill and knowledge will left unknown or lost.

This thesis reports the findings of research undergone for mapping expert to classification system. This research also interested to find out the possibility of using domain specific classification models that can contribute to more efficient retrieval of experts.

1.2. Motivation

1.2.1. Classifying Expert

Many organizations are emphasizing on expert finding due to expert discovery in an organization is important. The members of organization can share what expertise and knowledge they have at the moment. Thus, many organizations emphasize on identifying and finding experts of an organization with relevant expertise on a specific topic (Maryam K et al., 2009). In existing expert finding systems, profiles can be built from some sources, i.e. documents and emails, and used for expert recognition as the fundamental. Thus, expert classification is important for assisting on expert finding.

In an academic institute, finding expertise of researcher and specific information about academic staffs is important. For example, editors of conferences or journals usually need to find appropriate experts to review submitted papers (Kai-Hsiang Y et al., 2008). The expert discovery problem is solved by looking up expert-expertise databases. Therefore, one important approach for discover expert is classifying them. Classification is very important in many areas to assist in discovering information such as news, medical, book (digital library). Classification makes the study of a wide variety of aspect easy. Because of the usefulness of this classification, in this research, we are motivated to classify them by expertise.

1.2.2. Automated Classification

In general, classification can be conducted in automated or manually. One common approach used in automated classification is to use a “bag-of-words” representation. For training a classifier, the rate of occurrence of each word is used as a feature. In this thesis, we are motivated to exploit the widely available to build a category model in automated classification that can be contributed for learning these features based on the training texts indicated by both site. The category model will be used for categorizing new experts into right category based on the learnt features.

1.3. Objectives of the Research

According to the research motivation, the main objectives of this research are stated as follows:

1. To categorize expert into categories of domain specific classification system.

2. To compare the performance of two common term weighting algorithms, i.e. term-frequency (TF) and term-frequency – inverse document frequency (TF-IDF) in selecting feature for building category model of domain specific classification system.

3. To investigate the number of training texts required for building the category model.

1.4. Research Problem

This study focuses on the classification of expert into categories for domain specific classification system. The main goal is to categorize experts into categories based on their publication. However, there are some problems during the categorization. Firstly, an expert has much useful information for references. One of the useful information is bibliography. Over the time, the researcher will publish more and more publication. Since the detail of discussion for experts' bibliography may change over time, therefore the skill will remain unknown and the knowledge also will remain unknown. Thus, an automatic method of classification that categories expert into category would provide valuable insight into their expertise, knowledge, skill, etc.

Secondly, the number of training texts will affect the results that categorize experts into different categories. The size of training texts may change over time. Therefore, training of texts needs to adapt the change on size based on large training texts rather than small training texts. So, we need to find out a reasonable large number of training texts that can be used to build category model. The main reason for us to target large number of training texts is that the category model can be training will affect accuracy of expert categorized. Thirdly, term weighting will affect the result in feature selection. As such, we need to find out an approach that select important or useful term weighting especially in feature selection.

As the above mentioned, some problems related classification of expert into categories for domain specific classification system can be summarized as follows:

1. What is domain specific classification system can be used to categorize experts into category?

2. What is the numbers of training texts that can be used to build category model with reasonable accuracy?

3. How can we select relevant features from the training texts to contribute the category model?

1.5. Assumptions and Constraints

In general, classification evaluation is conducted using same training and testing texts, for example a collection of labelled of bibliography. However, in this thesis, our category model is trained using Call For Papers from conference site, whereas our model is tested using bibliography of experts from the same domain. Although our training and testing are collected from different data sets, we assume that they are similar in nature and characteristics because both texts are related publications. Another reason of using training data from WikiCFP rather than labelled bibliography is because the coverage of domain specific keywords is very high. The constraint of this research is number of categories evaluated may not cover all the categories for experts.

1.6. Scope of Research

This thesis focuses on the performance of expert categorized in expert classification system. This study adapts two classification systems which are related to the domain expertise of the expert. We have selected one classification system with categories that describes expertise in the broad manner and one classification system with categories that describes expertise in the specific manner. There are 30 categories selected for broad classification system (refer to broad classification model) from WikiCFP, whereas 2 main categories selected for specific classification system (refer to specific classification model) from ACM Taxonomy. For each category of broad classification system, the training texts are the CFP published by other authors in WikiCFP. Some specific words are collected for subcategories of specific classification system.

Two kinds of category model will be constructed, i.e. category model for broad classification and category model for specific category model. The study is designed to check whether our expert classification (ExpClass) method is able to classify expert into categories in Broad Classification System. Furthermore, the study is designed is to check whether Specific Classification System is suitable used in our ExpClass.

The study population is academic staff from USM. It has a total of 43 of experts. A total of 393 bibliographies are collected from PPSK USM Academic Staffs Bibliography.

1.7. Outline of Thesis

The reminder of this thesis is recognized as follows. Chapter 2 describes background and related work of this thesis. Chapter 3 describes the proposed framework that is used in this study. The work is evaluated in Chapter 4. The conclusion and some future directions are describes in Chapter 5.