A New Approach to Intranet Search Based on Information Extraction

Hang Li, Yunbo Cao
Microsoft Research Asia
5F Sigma Center
No.49 Zhichun Road,
Haidian, Beijing, China, 100080
{hangli, yucao}@microsoft.com

Jun Xu [*]
College of Software
Nankai University
No.94 Weijin Road,
Tianjin, China, 300071
nkxj@yahoo.com.cn

Yunhua Hu [*]
Dept. of Computer Science
Xi'an Jiaotong University
No 28, West Xianning Road,
Xi'an, China, 710049
yunhuahu@mail.xjtu.edu.cn

Shenjie Li [*]
Dept. of Computer Science
Hong Kong University of Science and Technology
Kowloon, Hong Kong, China
lisj@cs.ust.hk

Dmitriy Meyerzon
Microsoft Corporation
One Microsoft Way,
Redmond, WA, USA, 98052
dmitriym@microsoft.com

ABSTRACT

This paper is concerned with 'intranet search'. By intranet search, we mean searching for information on an intranet within an organization. We have found that search needs on an intranet can be categorized into types, through an analysis of survey results and an analysis of search log data. The types include searching for definitions, persons, experts, and homepages. Traditional information retrieval only focuses on search of relevant documents, but not on search of special types of information. We propose a new approach to intranet search in which we search for information in each of the special types, in addition to the traditional relevance search. Information extraction technologies can play key roles in such kind of 'search by type' approach, because we must first extract from the documents the necessary information in each type. We have developed an intranet search system called 'Information Desk'. In the system, we try to address the most important types of search first - finding term definitions, homepages of groups or topics, employees' personal information and experts on topics. For each type of search, we use information extraction technologies to extract, fuse, and summarize information in advance. The system is in operation on the intranet of Microsoft and receives accesses from about 500 employees per month. Feedbacks from users and system logs show that users consider the approach useful and the system can really help people to find information. This paper describes the architecture, features, component technologies, and evaluation results of the system.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process*; I.7.m [**Document and Text Processing**]: Miscellaneous

General Terms: Algorithms, Experimentation, Human Factors

Keywords: Intranet search, information extraction, metadata extraction, expert finding, definition search

1. INTRODUCTION

Internet search has made significant progress in recent years. In contrast, intranet search does not seem to be so successful. The IDC white paper entitled "The high cost of not finding information" [13] reports that information workers spend from 15% to 35% of their work time on searching for information and 40% of information workers complain that they cannot find the information they need to do their jobs on their company intranets.

Many commercial systems [35, 36, 37, 38, 39] have been developed for intranet search. However, most of them view intranet search as a problem of conventional relevance search. In relevance search, when a user types a query, the system returns a list of ranked documents with the most relevant documents on the top.

Relevance search can only serve *average* needs well. It cannot, however, help users to find information in a specific type, e.g., definitions of a term and experts on a topic. The characteristic of intranet search does not seem to be sufficiently leveraged in the commercial systems.

In this paper, we try to address intranet search in a novel approach. We assume that the needs of information access on intranets can be categorized into searches for information in different types. An analysis on search log data on the intranet of Microsoft and an analysis on the results of a survey conducted at Microsoft have verified the correctness of the assumption.

Our proposal then is to take a strategy of 'divide-and-conquer'. We first figure out the most important types of search, e.g., definition search, expert search. For each type, we employ information extraction technologies to extract, fuse, and summarize search results in advance. Finally, we combine all the types of searches together, including the traditional relevance

search, in a unified system. In this paper, we refer to the approach as 'search by type'. Search by type can also be viewed as a simplified version of Question Answering, adapted to intranet. The advantage of the new search approach lies in that it can help people find the types of information which relevance search cannot easily find. The approach is particularly reasonable on intranets, because in such space users are information workers and search needs are business oriented.

We have developed a system based on the approach, which is called 'Information Desk'. Information Desk can help users to find term definitions, homepages of groups or topics, employees' personal information and experts on topics, on their company intranets.

The system has been put into practical use since November 24th, 2004. Each month, about 500 Microsoft employees make access to the system. Both the results of an analysis on a survey and the results of an analysis on system log show that the features of definition search and homepage search are really helpful. The results also show that search by type is necessary at enterprise.

2. RELATED WORK

2.1 Intranet Search

The needs on search on intranets are huge. It is estimated that intranets at enterprises have tens or even hundreds of times larger data collections (both structured and unstructured) than internet. As explained above, however, many users are not satisfied with the current intranet search systems. How to help people access information on intranet is a big challenge in information retrieval. Much effort has been made recently on solutions both in industry and in academia.

Many commercial systems [35, 36, 37, 38, 39] dedicated to intranet search have been developed. Most of the systems view intranet search as a problem of conventional relevance search.

In the research community, ground designs, fundamental approaches, and evaluation methodologies on intranet search have been proposed.

Hawking et al [17] made ten suggestions on how to conduct high quality intranet search. Fagin et al [12] made a comparison between internet search and intranet search. Recently, Hawking [16] conducted a survey on previous work and made an analysis on the intranet search problem. Seven open problems on intranet search were raised in their paper.

Chen et al [3] developed a system named 'Cha-Cha', which can organize intranet search results in a novel way such that the underlying structure of the intranet is reflected. Fagin et al [12] proposed a new ranking method for intranet search, which combine various ranking heuristics. Mattox et al [25] and Craswell et al [7] addressed the issue of expert finding on a company intranet. They developed methods that can automatically identify experts in an area using documents on the intranet.

Stenmark [30] proposed a method for analyzing and evaluating intranet search tools.

2.2 Question Answering

Question Answering (QA) particularly that in TREC (http://trec.nist.gov/) is an application in which users type questions in natural language and the system returns short and usually single answers to the questions.

When the answer is a personal name, a time expression, or a place name, the QA task is called 'Factoid QA'. Many QA systems have been developed, [2, 4, 18, 20, 22, 27]. Factoid QA usually consists of the following steps: question type identification, question expansion, passage retrieval, answer ranking, and answer creation.

TREC also has a task of 'Definitional QA'. In the task, "what is <term>" and "who is <person>" questions are answered in a single combined text [1, 11, 15, 33, 34]. A typical system consists of question type identification, document retrieval, key sentence matching, kernel fact finding, kernel fact ranking, and answer generation.

3. OUR APPROACH TO INTRANET SEARCH

Search is nothing but collecting information based on users' information access requests. If we can correctly gather information on the basis of users' requests, then the problem is solved. Current intranet search is not designed along this direction. Relevance search can help create a list of ranked documents that serve only *average* needs well. The limitation of this approach is clear. That is, it cannot help users to find information of a specific type, e.g., definitions of a term. On the other hand, Question Answering (QA) is an ideal form for information access. When a user inputs a natural language question or a query (a combination of keywords) as a description of his search need, it is ideal to have the machine 'understand' the input and return only the necessary information based on the request. However, there are still lots of research work to do before putting QA into practical uses. In short term, we need consider adopting a different approach.

One question arises here: can we take a hybrid approach? Specifically, on one hand, we adopt the traditional approach for search, and on the other hand, we realize some of the most frequently asked types of search with QA. Finally, we integrate them in a single system. For the QA part, we can employ information extraction technologies to extract, fuse, and summarize the results in advance. This is exactly the proposal we make to intranet search.

Can we categorize users' search needs easily? We have found that we can create a hierarchy of search needs for intranet search, as will be explained in section 4.

On intranets, users are information workers and their motivations for conducting search are business oriented. We think, therefore, that our approach may be relatively easily realized on intranets first. (There is no reason why we cannot apply the same approach to the internet, however.)

To verify the correctness of the proposal, we have developed a system and made it available internally at Microsoft. The system called Information Desk is in operation on the intranet of Microsoft and receives accesses from about 500 employees per month.

At Information Desk, we try to solve the most important types of search first - find term definitions, homepages of groups or topics, experts on topics, and employees' personal information. We are

also trying to increase the number of search types, and integrate them with the conventional relevance search. We will explain the working of Information Desk in section 5.

4. ANALYSIS OF SEARCH NEEDS

In this section, we describe our analyses on intranet search needs using search query logs and survey results.

4.1 Categorization of Search Needs

In order to understand the underlying needs of search queries, we would need to ask the users about their search intentions. Obviously, this is not feasible. We conducted an analysis by using query log data. Here query log data means the records on queries typed by users, and documents clicked by the users after sending the queries.

Our work was inspirited by those of Rose and Levinson [28]. In their work, they categorized the search needs of users on *internet* by analyzing search query logs.

We tried to understand users' search needs on intranet by identifying and organizing a manageable number of categories of the needs. The categories encompass the majority of actual requests users may have when conducting search on an intranet.

We used a sample of queries from the search engine of the intranet of Microsoft. First, we brainstormed a number of categories, based on our own experiences and previous work. Then, we modified the categories, including adding, deleting, and merging categories, by assigning queries to the categories.

Given a query, we used the following information to deduce the underlying search need:

- the query itself
- the documents returned by the search engine
- the documents clicked on by the user

For example, if a user typed a keyword of '.net' and clicked a homepage of .net, then we judged that the user was looking for a homepage of .net.

As we repeated the process, we gradually reached the conclusion that search needs on intranet can be categorized as a hierarchical structure shown in Figure 1. In fact, the top level of the hierarchy resembles that in the taxonomy proposed by Rose and Levinson for internet [28]. However, the second level differs. On intranet, users' search needs are less diverse than those on internet, because the users are information workers and their motivations of conducting search are business oriented.

There is a special need called 'tell me about' here. It is similar to the traditional relevance search. Many search needs are by nature difficult to be categorized, for example, "I want to find documents related to both .net and SQL Server". We can put them into the category.

We think that the search needs are not Microsoft specific; one can image that similar needs exist in other companies as well.
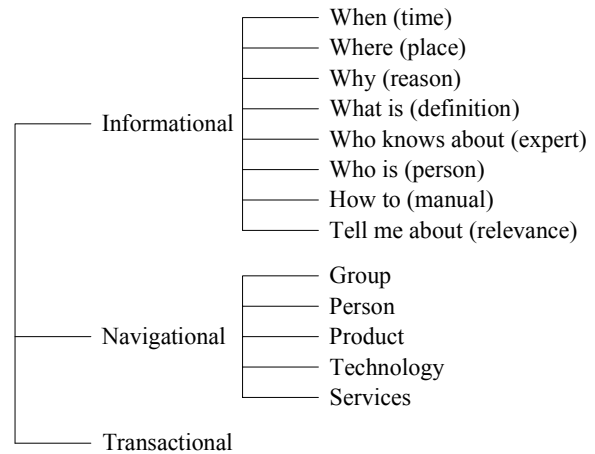

Figure 1. Categories of search needs

4.2 Analysis on Search Needs – by Query Log

We have randomly selected 200 unique queries and tried to assign the queries to the categories of search needs described above. Table 1 shows the distribution. We have also picked up the top 350 frequently submitted queries and assigned them to the categories. Table 2 shows the distribution. (There is no result for 'why', 'what is', and 'who knows about', because it is nearly impossible to guess users' search intensions by only looking at query logs.)

For random queries, informational needs are dominating. For high frequency queries, navigational needs are dominating. The most important types for random queries are relevance search, personal information search, and manual search. The most important types for high frequency queries are home page search and relevance search.

4.3 Analysis on Search Needs – by Survey

We can use query log data to analyze users' search needs, as described above. However, there are two shortcomings in the approach. First, sometimes it is difficult to guess the search intensions of users by only looking at query logs. This is especially true for the categories of 'why' and 'what'. Usually it is hard to distinguish them from 'relevance search'. Second, query log data cannot reveal users' potential search needs. For example, many employees report that they have needs of searching for experts on specific topics. However, it is difficult to find expert searches from query log at a conventional search engine, because users understand that such search is not supported and they do not conduct the search.

To alleviate the negative effect, we have conducted another analysis through a survey. Although a survey also has limitation (i.e., it only asks people to answer pre-defined questions and thus can be biased), it can help to understand the problem from a different perspective.

Table 1. Distribution of search needs for random queries

Category of Search Needs	Percentage
When	0.02
Where	0.02
Why	NA
What is	NA
Who knows about	NA
Who is	0.23
How to	0.105
Tell me about	0.46
Informational total	**0.835**
Groups	0.03
Persons	0.005
Products	0.02
Technologies	0.02
Services	0.06
Navigational total	**0.135**
Transactional	**0.025**
Other	**0.005**

Table 2. Distribution of search needs for high frequency queries

Category of Search Needs	Relative Prevalence
When	0.0057
Where	0.0143
Why	NA
What is	NA
Who knows about	NA
Who is	0.0314
How to	0.0429
Tell me about	0.2143
Informational total	**0.3086**
Groups	0.0571
Persons	0.0057
Products	0.26
Technologies	0.0829
Services	0.2371
Navigational total	**0.6428**
Transactional	**0.0086**
Other	**0.04**

I have experiences of conducting search at Microsoft intranet to look for the web sites (or homepages) of (multiple choice)
- technologies 74 %
- products 74 %
- services 68 %
- projects 68 %
- groups 60 %
- persons 42 %
- none of the above 11 %

I have experiences of conducting search at Microsoft intranet in which the needs can be translated into questions like? (multiple choice)
- 'what is' - e.g., "what is blaster" 77 %
- 'how to' - "how to submit expense report" 54 %
- 'where' - e.g., "where is the company store" 51 %
- 'who knows about' - e.g., "who knows about data mining" 51 %
- 'who is' - e.g., "who is Rick Rashid" 45 %
- 'when' - e.g., "when is TechFest'05 " 42 %
- 'why' - e.g., "why do Windows NT device drivers contain trusted code" 28 %
- none of the above 14 %

I have experiences of conducting search at Microsoft intranet in order to (multiple choice)
- download a software, a document, or a picture. E.g., "getting MSN logo" 71 %
- make use of a service. E.g., "getting a serial number of Windows" 53 %
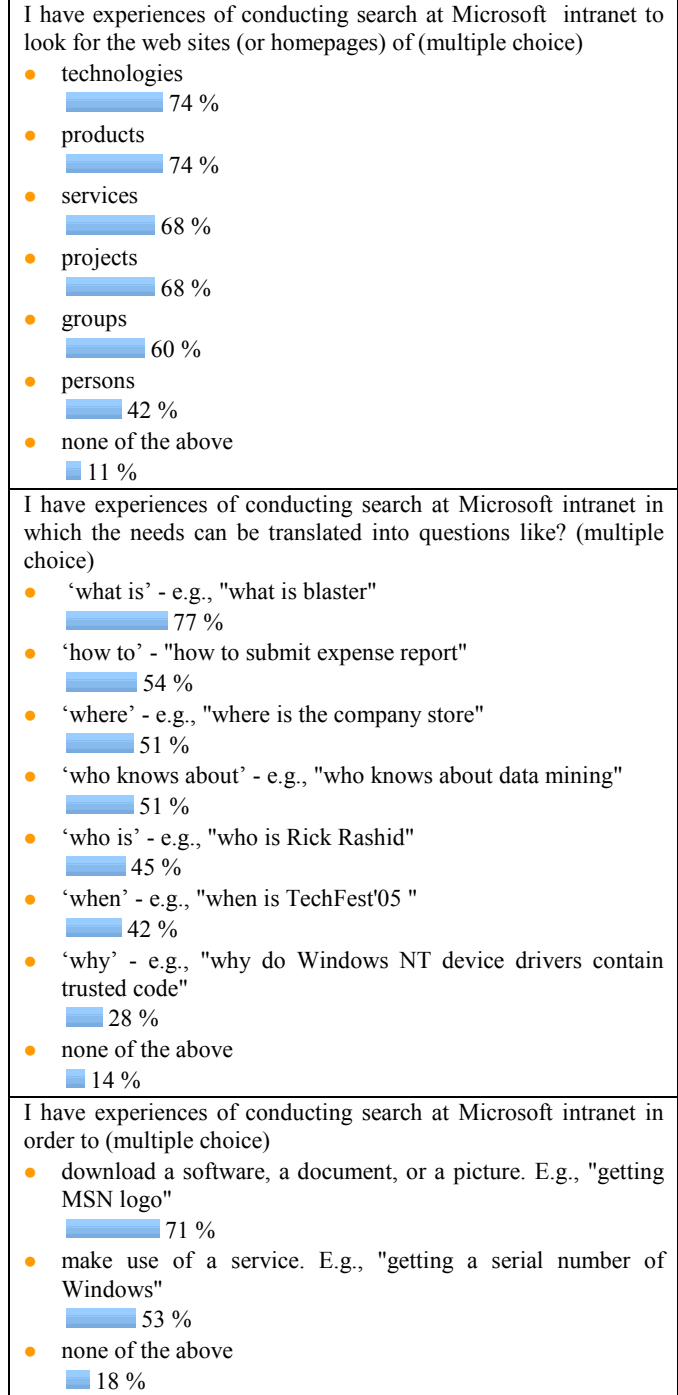- none of the above 18 %

Figure 2. Survey results on search needs

In the survey, we have asked questions regarding to search needs at enterprise. 35 Microsoft employees have taken part in the survey. Figure 2 shows the questions and the corresponding results.

We see from the answers that definition search, manual search, expert finding, personal information search, and time schedule search are requested by the users. Homepage finding on technologies and products are important as well. Search for a download site is also a common request.
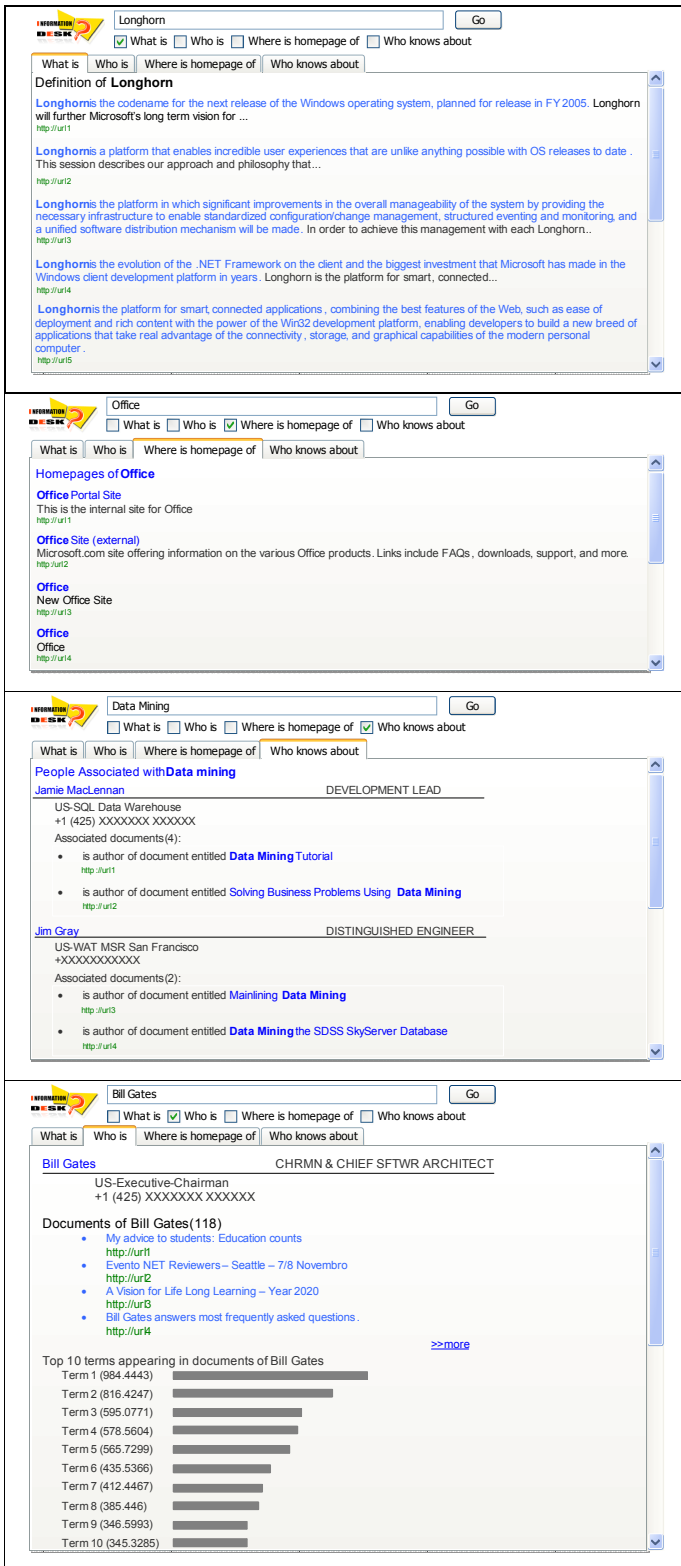
Figure 3: Information Desk system

5. INFORMATION DESK

5.1 Features

Currently Information Desk provides four types of search. The four types are:

1. 'what is' – search of definitions and acronyms. Given a term, it returns a list of definitions of the term. Given an acronym, it returns a list of possible expansions of the acronym.

2. 'who is' – search of employees' personal information. Given the name of a person, it returns his/her profile information, authored documents and associated key terms.

3. 'where is homepage of' – search of homepages. Given the name of a group, a product, or a technology, it returns a list of its related home pages.

4. 'who knows about' – search of experts. Given a term on a technology or a product, it returns a list of persons who might be experts on the technology or the product.
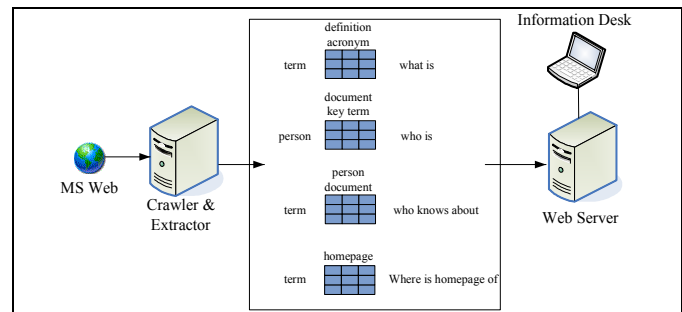

Figure 4. Workflow of Information Desk

There are check boxes on the UI, and each represents one search type. In search, users can designate search types by checking the corresponding boxes and then submit queries. By default, all the boxes are checked.

For example, when users type 'longhorn' with the 'what is' box checked, they get a list of definitions of 'Longhorn' (the first snapshot in figure 3). Users can also search for homepages (team web sites) related to 'Office', using the 'where is homepage' feature (the second snapshot in figure 3). Users can search for experts on, for example, 'data mining' by asking 'who knows about data mining' (the third snapshot in figure 3). Users can also get a list of documents that are automatically identified as being authored by 'Bill Gates', for example, with the 'who is' feature (the last snapshot in figure 3). The top ten key terms found in his documents are also given.

Links to the original documents, from which the information has been extracted, are also available on the search result UIs.

5.2 Technologies

5.2.1 Architecture

Information Desk makes use of information extraction technologies to support the search by type feaatures. The technologies include automatically extracting document metadata and domain specific knowledge from a web site using information extraction technologies. The domain specific knowledge includes definition, acronym, and expert. The document metadata includes title, author, key term, homepage. Documents are in the form of Word, PowerPoint, or HTML. Information Desk stores all the data in Microsoft SQL Server and provides search using web

services. Figure 4 shows the workflow of Information Desk. Currently, there are 4 million documents crawled from the Microsoft intranet.

Below we explain each feature in details. Table 3 shows which feature employs what kind of mining technology.

Table 3. Information extraction technologies employed

	'What is'	'Who is'	'Who knows about'	'Where is homepage'
Definition extraction	Yes			
Acronym extraction	Yes			
Homepage finding				Yes
Title extraction		Yes	Yes	
Author extraction		Yes	Yes	
Key term extraction		Yes	Yes	
Expert mining			Yes	

5.2.2 'What is'

There are two parts in the feature: definition finding and acronym recognition.

In definition finding, we extract from the entire collection of documents <term, definition, score> triples. They are respectively a term, a definitional excerpt of the term, and its score representing its likelihood of being a good definition. We assign the scores using a statistical model. Both paragraphs and sentences can be considered as definition excerpts in our approach. Currently, we only consider the use of paragraphs.

As model, we employ SVM (Support Vector Machines) [31], which identifies whether a given paragraph is a definition of the first noun phrase (term) in the paragraph. There are positive features in the SVM model. For example, if the term appears at the beginning of the paragraph or repeatedly occurs in the paragraph, then it is likely the paragraph is a (good) definition on the term. There are also negative features. If words like 'she', 'he', or 'said' occurs in the paragraph, or many adjectives occur in the paragraph, then it is likely the paragraph is not a (good) definition.

In search, given a query term, we retrieve all the triples matched against the query term and present the corresponding definitions in descending order of the scores.

The top 1 and top 3 precision of our approach in definition ranking are 0.550 and 0.887 respectively. They are much better than the baseline method of employing relevance search.

Methods for extracting definitions from documents have been proposed [1, 10, 11, 15, 21, 24, 33]. All of the methods resorted to human-defined rules for the extraction and did not consider ranking of definitions. In Information Desk, we rank definitions according to their likelihoods of being good definitions, represented by SVM scores. See [32] for details.

In acronym recognition, we find candidate acronym and candidate expansion pairs from text using pattern matching. There are ten types of patterns. For example, one of them is '<expansion> (<acronym>)' in which <expansion> denotes a phrase with the

first letters in the words capitalized and <acronym> denotes a sequence of the capitalized letters in the same order. The pattern matches sentences such as "Active Directory is implemented using the Lightweight Directory Access Protocol (LDAP)". We then store all the acronyms, their expansions, and the numbers of occurrences of the expansions.

In search, given an acronym, we retrieve all the expansions against the acronym and present the corresponding expansions in descending order of their numbers of occurrences.

5.2.3 'Who is'

We first harvest all the employees' personal information from a database. It includes name, alias, title, and contact information. We next automatically extract titles and authors from all the Word and PowerPoint documents on the intranet. With the extracted titles and authors we bring together all the documents to each person, which are thought authored by him/her. Finally, we extract key terms from the documents for each person and pick up the top ten key terms in terms of TF-IDF. This feature lies mainly on document metadata extraction.

Metadata of documents such as title and author is useful for document processing. However, people seldom define document metadata by themselves. We collected 6,000 Word and 6,000 PowerPoint documents and examined how many titles and authors in the file properties are correct. We found that the accuracies were only 0.265 and 0.126 respectively.

We take a machine learning approach to automatically extract titles and authors from the bodies of Office documents, as shown in Figure 5. We annotate titles in sample documents (for Word and PowerPoint respectively) and take them as training data, train statistical models, and perform title extraction using the trained models. In the models, we mainly utilize format information such as font size as features. As models, we employ Perceptron with Uneven Margins [23].

Experimental results indicate that our approach works well for title extraction from general documents. Our method can significantly outperform the baselines: one that always uses the first lines as titles and the other that always uses the lines in the largest font sizes as titles. Precision and recall for title extraction from Word are 0.875 and 0.899 respectively, and precision and
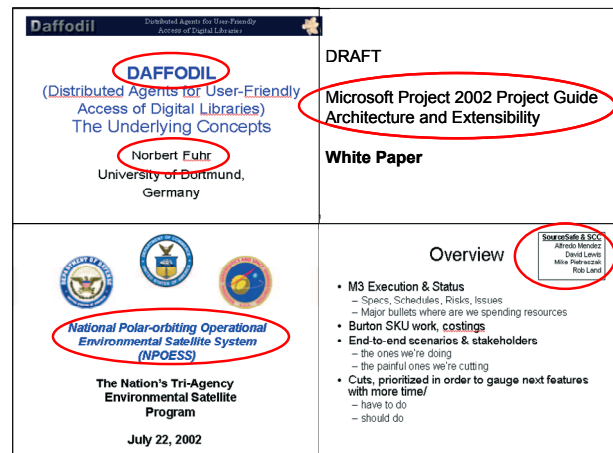

Figure 5. Title and author extraction from four example PowerPoint documents

recall for title extraction from PowerPoint are 0.907 and 0.951 respectively.

Metadata extraction has been intensively studied. For instance, Han et al [14] proposed a method for metadata extraction from research papers. They considered the problem as that of classification based on SVM. They mainly used linguistic information as features. To the best of our knowledge, no previous work has been done on metadata extraction from general documents. We report our title extraction work in details in [19].

The feature of 'who is' can help find documents authored by a person, but existing in different team web sites. Information extraction (specifically metadata extraction) makes the aggregation of information possible.

5.2.4 'Who knows about'

The basic idea for the feature is that if a person has authored many documents on an issue (term), then it is very likely that he/she is an expert on the issue, or if the person's name co-occurs in many times with the issue, then it is likely that he/she is an expert on the issue.

As described above, we can extract titles, authors, and key terms from all the documents. In this way, we know how many times each person is associated with each topic in the extracted titles and in the extracted key terms. We also go through all the documents and see how many times each person's name co-occurs with each topic in text segments within a pre-determined window size.

In search, we use the three types of information: topic in title, topic in key term, and topic in text segment to rank persons, five persons for each type. We rank persons with a heuristic method and return the list of ranked persons. A person who has several documents with titles containing the topic will be ranked higher than a person whose name co-occurs with the topic in many documents.

It appears that the results of the feature largely depend on the size of document collection we crawl. Users' feedbacks on the results show that sometimes the results are very accurate, however, sometimes they are not (due to the lack of information).

Craswell et al. developed a system called 'P@NOPTIC', which can automatically find experts using documents on an intranet [7]. The system took documents as plain texts and did not utilize metadata of documents as we do at Information Desk.

5.2.5 'Where is homepage of'

We identify homepages (team web sites) using several rules. Most of the homepages at the intranet of Microsoft are created by SharePoint, a product of Microsoft. From SharePoint, we can obtain a property of each page called 'ContentClass'. It tells exactly whether a web page corresponds to a homepage or a team site. So we know it is a homepage (obviously, this does not apply in general). Next we use several patterns to pull out titles from the homepages. The precision of home page identification is nearly 100%.

In search, we rank the discovered home pages related to a query term using the URL lengths of the home pages. A home page with a shorter URL will be ranked higher.

TREC has a task called 'home/named page finding' [8, 9], which is to find home pages talking about a topic. Many methods have been developed for pursuing the task [5, 6, 26, 29]. Since we can identify homepages by using special properties on our domain, we do not consider employing a similar method.

6. EVALUATION

Usually it is hard to conduct evaluation on a practical system. We evaluated the usefulness of Information Desk by conducting a survey and by recording system logs.

We have found from analysis results that the 'what is' and 'where is homepage of' features are very useful. The 'who is' feature works well, but the 'who knows about' feature still needs improvements.

6.1 Survey Result Analysis

The survey described in section 4.3 also includes feedbacks on Information Desk.

Figure 6 shows a question on the usefulness of the features and a summary on the answers. We see that the features 'where is homepage of' and 'what is' are regarded useful by the responders in the survey.

Figure 7 shows a question on new features and a summary on the answers. We see that the users want to use the features of 'how to', 'when', 'where' and 'why' in the future. This also justifies the correctness of our claim on intranet search made in section 4.

Figure 8 shows a question on purposes of use and a digest on the results. About 50% of the responders really want to use Information Desk to search for information.

There is also an open-ended question asking people to make comments freely. Figure 9 gives some typical answers from the responders. The first and second answers are very positive, while the third and fourth point out the necessity of increasing the coverage of the system.
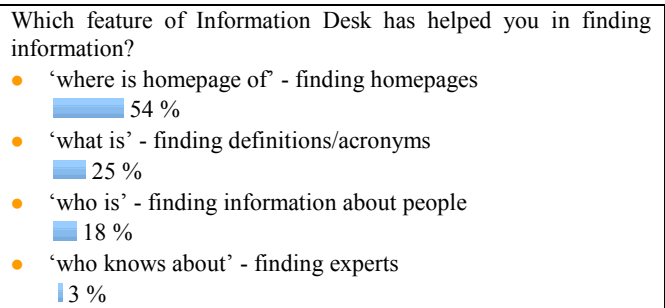
Which feature of Information Desk has helped you in finding information?
- 'where is homepage of' - finding homepages
  54 %
- 'what is' - finding definitions/acronyms
  25 %
- 'who is' - finding information about people
  18 %
- 'who knows about' - finding experts
  3 %

Figure 6. Users' evaluation on Information Desk

What kind of new feature do you want to use at Information Desk? (multiple choice)
- 'how to' - e.g., "how to activate Windows"
  57 %
- 'when' - e.g., "when is Yukon RTM"
  57 %
- 'where' - e.g., "where can I find an ATM"
  39 %
- 'why' - e.g., "why doesn't my printer work"
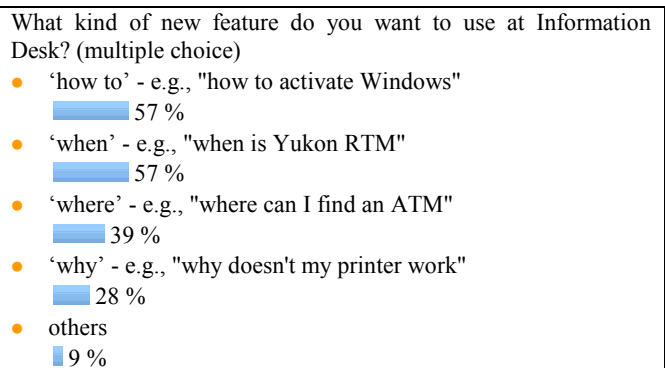 28 %
- others
  9 %

Figure 7. New features expected by users

I visited Information Desk today to
- conduct testing on Information Desk
  54 %
- search for information related to my work
  46 %

Figure 8. Motivation of using Information Desk

Please provide any additional comments, thanks!
- This is a terrific tool! Including 'how to' and 'when' capabilities will put this in the 'can't live without it' category.
- Extremely successful searching so far! Very nice product with great potential.
- I would like to see more 'Microsoftese' definitions. There is a lot of cultural/tribal knowledge here that is not explained anywhere.
- Typing in my team our website doesn't come up in the results, is there any way we can provide content for the search tool e.g., out group sharepoint URL?
- …

Figure 9. Typical user comments to Information Desk

6.2 System Log Analysis

We have made log during the running of Information Desk. The log includes user IP addresses, queries and clicked documents (recall that links to the original documents, from which information has been extraction, are given in search). The log data was collected from 1,303 unique users during the period from November 26th, 2004 to February 22nd, 2005. The users were Microsoft employees.

In the log, there are 9,076 query submission records. The records include 4,384 unique query terms. About 40% of the queries are related to the 'what is' feature, 29% related to 'where is homepage of', 30% related to 'who knows about' and 22% related to 'who is'. A query can be related to more than one feature.

In the log, there are 2,316 clicks on documents after query submissions. The numbers of clicks for the 'what is', 'where is homepage of', 'who knows about', and 'who is' features are 694, 1041, 200 and 372, respectively. Note that for 'what is', 'where is home page of', and 'who knows about' we conduct ranking on retrieved information. The top ranked results are considered to be the best. If a user has clicked a top ranked document, then it means that he is interested in the document, and thus it is very likely he has found the information he looks for. Thus a system which has higher average rank of clicks is better than the other that does not. We used average rank of clicked documents to evaluate the performances of the features. The average ranks of clicks for 'what is', 'where is homepage of' and 'who knows about' are 2.4, 1.4 and 4.7 respectively. The results indicate that for the first two features, users usually can find information they look for on the top three answers. Thus it seems safe to say that the system have achieved practically acceptable performances for the two features. As for 'who is', ranking of a person's documents does not seem to be necessary and the performance should be evaluated in a different way. (For example, precision and recall of metadata extraction as we have already reported in section 5).

7. CONCLUSION

In this paper, we have investigated the problem of intranet search using information extraction.

- Through an analysis of survey results and an analysis of search log data, we have found that search needs on intranet can be categorized into a hierarchy.

- Based on the finding, we propose a new approach to intranet search in which we conduct search for each special type of information.

- We have developed a system called 'Information Desk', based on the idea. In Information Desk, we provide search on four types of information - finding term definitions, homepages of groups or topics, employees' personal information and experts on topics. Information Desk has been deployed to the intranet of Microsoft and has received accesses from about 500 employees per month. Feedbacks from users show that the proposed approach is effective and the system can really help employees to find information.

- For each type of search, information extraction technologies have been used to extract, fuse, and summarize information in advance. High performance component technologies for the mining have been developed.

As future work, we plan to increase the number of search types and combine them with conventional relevance search.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] S. Blair-Goldensohn, K.R. McKeown, A.H. Schlaikjer. A Hybrid Approach for QA Track Definitional Questions. *In Proc. of Twelfth Annual Text Retrieval Conference (TREC-12),* NIST, Nov., 2003.

[2] E. Brill, S. Dumais, and M. Banko, *An Analysis of the AskMSR Question-Answering System*, EMNLP 2002

[3] M. Chen, A. Hearst, A. Marti, J. Hong, and J. Lin, Cha-Cha: A System for Organizing Intranet Results. *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems. Boulder, CO*. Oct. 1999.

[4] C. L. A. Clarke, G. V. Cormack, T. R. Lynam, C. M. Li, and G. L. McLearn, *Web Reinforced Question Answering (MultiText Experiments for TREC 2001)*. TREC 2001

[5] N. Craswell, D. Hawking, and S.E. Robertson. Effective site finding using link anchor information. *In Proc. of the 24th annual international ACM SIGIR conference on research and development in information retrieval,* pages 250--257, 2001.

[6] N. Craswell, D. Hawking, and T. Upstill. TREC12 Web and Interactive Tracks at CSIRO. *In TREC12 Proceedings*, 2004.

[7] N. Craswell, D. Hawking, A. M. Vercoustre, and P. Wilkins. P@noptic expert: Searching for experts not just for documents. *Poster Proceedings of AusWeb'01*,

2001b./urlausweb.scu.edu.au/aw01/papers/edited/vercoustre/paper.htm.

[8] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. *Overview of the TREC-2003 Web Track. In NIST Special Publication: 500-255, The Twelfth Text REtrieval Conference (TREC 2003)*, Gaithersburg, MD, 2003.

[9] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. *Task Descriptions: Web Track 2003. In TREC12 Proceedings,* 2004.

[10] H. Cui, M-Y. Kan, and T-S. Chua. Unsupervised Learning of Soft Patterns for Definitional Question Answering, *Proceedings of the Thirteenth World Wide Web conference (WWW 2004),* New York, May 17-22, 2004.

[11] A. Echihabi, U.Hermjakob, E. Hovy, D. Marcu, E. Melz, D. Ravichandran. Multiple-Engine Question Answering in TextMap. *In Proc. of Twelfth Annual Text Retrieval Conference (TREC-12)*, NIST, Nov., 2003.

[12] R. Fagin, R. Kumar, K. S. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin, and D. P. Williamson. Searching the workplace web. *Proc. 12th World Wide Web Conference*, Budapest, 2003.

[13] S. Feldman and C. Sherman. The high cost of not finding information. *Technical Report #29127, IDC, April 2003.*

[14] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic Document Metadata Extraction using Support Vector Machines. *In Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries,* 2003

[15] S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, J. Bensley. Answer Mining by Combining Extraction Techniques with Abductive Reasoning. *In Proc. of Twelfth Annual Text Retrieval Conference (TREC-12),* NIST, Nov., 2003.

[16] D. Hawking. Challenges in Intranet search. *Proceedings of the fifteenth conference on Australasian database*. Dunedin, New Zealand, 2004.

[17] D. Hawking, N. Craswell, F. Crimmins, and T. Upstill. Intranet search: What works and what doesn't. *Proceedings of the Infonortics Search Engines Meeting*, San Francisco, April 2002.

[18] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C. Y. Lin. *Question Answering in Webclopedia*. TREC 2000

[19] Y. Hu, H. Li, Y. Cao, D. Meyerzon, and Q. Zheng. Automatic Extraction of Titles from General Documents using Machine Learning. To appear at Proc. of Joint Conference on Digital Libraries (JCDL), 2005. Denver, Colorado, USA. 2005.

[20] A. Ittycheriah and S. Roukos, *IBM's Statistical Question Answering System-TREC 11*. TREC 2002

[21] J. Klavans and S. Muresan. DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text. *In Proceedings of AMIA Symposium 2000.*

[22] C. C. T. Kwok, O. Etzioni, and D. S. Weld, *Scaling question answering to the Web*. WWW-2001: 150-161

[23] Y. Li, H Zaragoza, R Herbrich, J Shawe-Taylor, and J. S. Kandola. The Perceptron Algorithm with Uneven Margins. in *Proceedings of ICML'02*.

[24] B. Liu, C. W. Chin, and H. T. Ng. Mining Topic-Specific Concepts and Definitions on the Web. *In Proceedings of the twelfth international World Wide Web conference (WWW-2003),* 20-24 May 2003, Budapest, HUNGARY.

[25] D. Mattox, M. Maybury and D. Morey. Enterprise Expert and Knowledge Discovery. *Proceedings of the HCI International '99 (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Communication, Cooperation, and Application Design-Volume 2 - Volume 2*. 1999.

[26] P. Ogilvie and J. Callan. Combining Structural Information and the Use of Priors in Mixed Named-Page and Homepage Finding. *In TREC12 Proceedings*, 2004.

[27] D. R. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. *Probabilistic question answering on the web*. WWW 2002: 408-419

[28] D. E. Rose and D. Levinson. Understanding user goals in web search. *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, 2004 New York, USA.

[29] J. Savoy, Y. Rasolofo, and L. Perret, L. Report on the TREC-2003 Experiment: Genomic and Web Searches. *In TREC12 Proceedings,* 2004.

[30] D. Stenmark. A Methodology for Intranet Search Engine Evaluations. *Proceedings of IRIS22, Department of CS/IS*, University of Jyväskylä, Finland, August 1999.

[31] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[32] J. Xu, Y. Cao, H. Li, and M. Zhao. Ranking Definitions with Supervised Learning Methods. In Proc. of 14th International World Wide Web Conference (WWW05), Industrial and Practical Experience Track, Chiba, Japan, pp.811-819, 2005.

[33] J. Xu, A. Licuanan, R. Weischedel. TREC 2003 QA at BBN: Answering Definitional Questions. *In Proc. of 12th Annual Text Retrieval Conference (TREC-12),* NIST, Nov., 2003.

[34] H. Yang, H. Cui, M. Maslennikov, L. Qiu, M-Y. Kan, and T-S. Chua, *QUALIFIER in TREC-12 QA Main Task*. TREC 2003: 480-488

[35] Intellectual capital management products. *Verity, http://www.verity.com/*

[36] IDOL server. *Autonomy, http://www.autonomy.com/content/home/*

[37] Fast data search. *Fast Search & Transfer, http://www.fastsearch.com/*

[38] Atomz intranet search. *Atomz, http://www.atomz.com/*

[39] Google Search Appliance. *Google, http://www.google.com/enterprise/*