Task-Centric Document Recommendation via Context-Specific Terms

Surendra Sarnikar, J. Leon Zhao

Department of MIS
Eller College of Management
University of Arizona, Tucson, AZ 85721
{sarnikar, lzhao}@eller.arizona.edu

Abstract

Context-specific document recommender systems rely on the accurate identification of context descriptors from unstructured textual information to identify highly relevant documents. In this paper, we propose two term-weighting measures, "normal distance" and "adjusted inverse polysemy", to enable the retrieval of relevant documents with higher precision. We analyze the performance of the proposed measures and present results with respect to a domain-specific corpus.

Introduction

Document recommender systems can be used to recommend task specific documents by matching contextual information such as email messages or task descriptions, with documents from a knowledge base (Abecker et al., 2000; Kwan and Balasubramanian, 2003; Sarnikar and Zhao, 2005).

Task-centric document recommendation involves the proactive recommendation of task-specific documents to a user, within the context of the task. Although traditional information retrieval and filtering techniques can be used to recommend documents with the task, the conditions under which task-specific document recommenders are used is significantly different from that of traditional information retrieval systems, such as search engines.

While conventional information retrieval systems are typically initiated by a user and function using specific keywords that are supplied by the user, task-centric document recommender systems rely on contextual information to proactively deliver relevant documents. In addition, a task specific document recommender is typically embedded within the application that supports a user's work process, such as workflow systems, collaboration systems or an email application. Hence, the task-specific document recommenders need to be adapted to deliver few but highly relevant documents.

In order to reduce the user cognitive load and reduce information overload, the task-specific recommender systems need to be optimized to perform at high precision at the expense of recall. However, in order to retrieve highly relevant documents from a knowledge-base, it is important to identify key terms from the contextual information that describes the knowledge requirement with highest specificity. This information can in turn be used to develop weighted queries that are better able to describe the knowledge requirements of the task.

Differential weighting of query terms based on their importance increases the precision of a retrieval system by enabling early detection of the relevant documents. In this paper, we propose two term weighting measures for differentiating key terms from generic terms in contextual information.

The proposed measures rely on two external sources of information to estimate the weights of the query terms; a thesaurus to determine the polysemy counts of the terms in the query, and a generic corpus to identify the background occurrence frequency of a term. We utilize the difference in occurrence frequency of a term in a generic and a domain specific corpus, and its polysemy count to determine the relative importance of a query term within a query.

In the remaining sections, we analyze the effect of the proposed term-weighting mechanisms on precision and recall and compare it with the performance of a simple retrieval system without term weighting. We begin by briefly reviewing previous work in this area.

Previous Work

Several systems have been proposed for retrieving documents based on context. In order to recommend relevant documents, the proposed mechanisms rely on relevance feedback (Anick and Tipirneni, 1999), query expansion using an external thesaurus (Liu and Chu, 2005) or on heuristic algorithms that analyze user behavior (Budzik and Hammond, 2000). Finkelstein et al., (2002) describe a semantic analysis-based context specific search assistant that outperforms all the popular search engines in retrieval precision when comparing the first ten retrieved documents. However, most of the proposed systems require the user to select the search terms or do not differentiate between key terms and generic terms from unstructured contextual information.

Differential weighting of query terms forms the basis of modern information retrieval (Jones 1972; van Rijsbergen, 1975; Salton and Buckley 1988). The most widely used and robust mechanism for term weighting is based on document frequency and term frequency, or the tf*idf measure. The tf*idf measure is based on the assumption that highly specific terms are used less frequently than generic terms (Sparck Jones, 1972).

However, in the case of domain specific copora consisting of short length documents such as article abstracts, this assumption is weakened. Given the short

length of documents, most terms occur only once in a short document. In addition, domain terms are more frequently used and the use of generic terms is minimized to reduce ambiguity. Hence an additional term-weighting mechanism is required to identify the domain specific terms and to compensate for the above deviations from the assumptions underlying the tf*idf measure. In the following sections, we present further evidence of this deviation and its effect on retrieval performance.

Data

We use the CACM corpus to analyze the effectiveness of the proposed term weighting mechanisms. The CACM corpus consists of 3204 documents. The documents consist of titles and abstracts of articles published in the Communications of the ACM journal. The articles relate to a wide variety of topics in the computer science field. The CACM document set also includes 64 queries and a list of documents judged relevant to the queries. The queries describe the knowledge requirements of computer science researchers. For example, query 4 from the CACM corpus is given below:

> I'm interested in mechanisms for communicating between disjoint processes, possibly, but not exclusively, in a distributed environment. I would rather see descriptions of complete mechanisms, with or without implementations, as opposed to theoretical work on the abstract problem. Remote procedure calls and message-passing are examples of my interests.

Retrieval Analysis

We used a simple best match retrieval system to retrieve documents using the given queries. We used a widely available Apache Lucene retrieval engine to retrieve the documents. The default similarity measure provided with the search engine was used for retrieval (Lucene API, 2004). The similarity measure, which is a variation of the cosine measure, is as follows:

$$Sim(q,d) = \sum_{t \in q} \frac{tf_{t,q} idf_t}{norm_q} . \frac{tf_{t,d} . idf_t}{norm_d} . coord_{q,d} . weight_t$$

Where,

Sim(q, d) is the level of similarity between the query *q* and document *d.*

idf_t is the inverse document frequency of the term *t*.

$tf_{t,d}$ is the frequency of the term *t* within document *d*.

$norm_q$ is the normalization factor for the query and is given by $\sqrt{\sum_{t \in q} (idf_t)^2}$

$norm_d$ is the normalization factor for document length

$coord_{q,d}$ is the overlap in terms of query q and document d

$weight_t$ is the weight of the term *t* in query *q*

$tf_{t,q,}$ the frequency of the term *t* in query *q* is set to 1 in all the runs.

In the initial run, all the terms in the query were assigned equal weights. The precision and recall curve for the initial run is given in Figure 1.

In order to identify the factors affecting the performance of the retrieval system, we analyzed the queries and the corresponding documents retrieved by the system. We examined each query and identified the domain specific terms and generic terms present in the query. For example, the query 4 given above can be categorized into subsets of domain specific terms {communication, disjoint, process, distributed}, {remote procedure call}, {message-passing} and generic terms such as {possibly, interested, descriptions, rather, examples, problem}. While the domain-specific terms are indicative of the knowledge requirements and are most useful in identifying relevant documents when using best-match retrieval systems, the contribution of the generic terms to the query precision is minimal or negative.

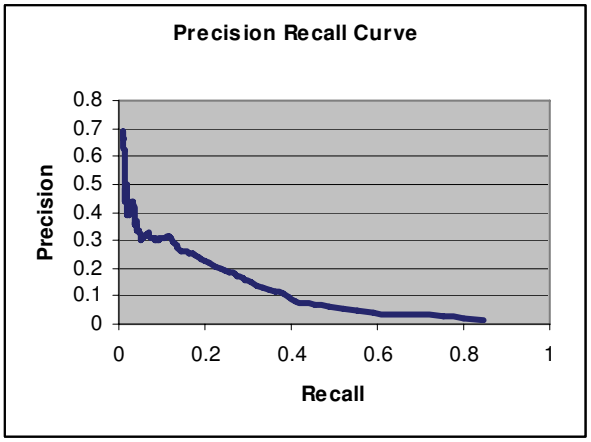

Figure 1. Precision Recall Curve for Initial Run

The documents retrieved by the retrieval systems can be categorized as those matching a subset of query terms that contain mostly domain specific terms or as those matching mostly generic terms. For example, document 2939 was retrieved in response to query 4 as the top ranked document as it contained a large number of generic terms from the query such as *abstract, complete, example, implement, mechanism, procedure, process,* and *work*. However, the document did not contain any of the subsets of domain specific query terms identified above and was not relevant to the query. On the other hand, a relevant document, containing the query terms (communication, distributed, process) was given a lower rank as it contained only a few of the query terms.

On further analysis, we observed that some documents that contained all of the domain specific terms were also not relevant due to the polysemy problem. We further observed that documents that contained generic terms or only a part of the subset of domain specific terms were sometimes found to be relevant as they contained terms synonymous to the domain specific query terms. For example, several of the documents that were found to be relevant to query 4 on "communication between disjoint processes" contained only generic terms or only part of the subset of the domain specific terms but also contained terms synonymous or related to the domain specific query terms such as concurrent processes and synchronization.

Based on the above analysis, the documents retrieved by the best match retrieval system can be categorized as (1) Relevant documents retrieved by matching domain specific query terms, (2) Documents matching domain specific query terms but that are not relevant due to polysemy problem, (3) Documents that match generic query terms but are relevant as they contain term synonymous to domain specific query terms and (4) Documents that match generic query terms and are not relevant. A hierarchical classification of the retrieved documents is shown in Figure 2.

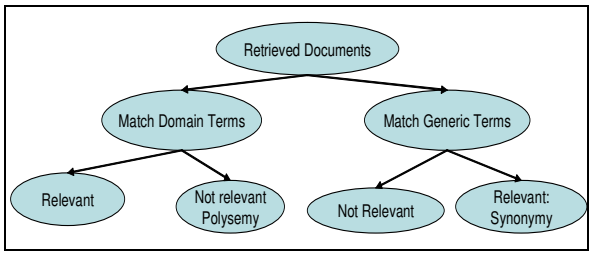

Figure 2. Categorization of Retrieved Documents

Context Specific Document Recommenders

A context-specific document recommender system retrieves documents from a knowledge repository that are relevant to the task-at-hand. A content-based context-specific document recommender can be embedded in email systems, workflow system and retrieve relevant documents by matching contextual information from task descriptions or email messages. However, unlike a search engine, which can be used for exploratory search, a context-specific document recommender system needs to recommend a few but highly relevant documents, thereby preventing information overload.

In this paper, our objective is to automatically identify the domain specific terms in a query, in order to give higher relevance weights to documents that match domain terms and lower relevance values to documents that match only the generic terms or match only some of the domain specific terms in a query. We hypothesize that as a result of increasing the weights of domain specific terms, documents that match the domain specific terms will be retrieved at higher threshold values, hence increasing the precision and recall at higher threshold values. Similarly, the documents that match mostly generic terms will be retrieved at lower threshold values. An overview of the term weighting process is shown in Figure 3.

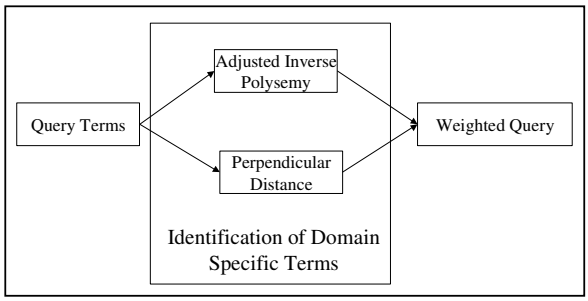

Figure 3. Term-weighting process

As a result of this re-arrangement of the domain term matching and generic term matching documents along the precision recall curve, the shape of the curve is altered to reflect higher precision and recall at higher threshold values and lower precision and recall at lower threshold values when compared with a best match retrieval without term-weighting. In-effect, assigning higher weights to domain specific terms retrieves a larger number of relevant documents at higher thresholds when compared to best-match retrieval without term weighting. In the following section we discuss the identification of domain specific terms and describe their characteristics

Identification of Domain Specific Terms

Several methods have been proposed for identifying domain specific terms using statistical measures that compare term document frequencies in generic and domain specific corpora (Vogel, 2003; Navigli et al., 2003). In this paper, we extend Vogel's (2003) method for identifying domain specific terms and extend it for use in document recommender systems.

We analyzed the occurrence of query terms in the domain specific CACM corpus and in a large generic corpus to identify the characteristics of domain specific terminology. We estimated a term's document frequency in a generic corpus by querying for its document frequency in Google's web document database. A scatter plot of the query terms along the dimensions of their inverse document frequency in a generic and domain specific corpus is given in Figure 4. The 45° line represents the region where the occurrence frequency of a term is similar in both the domain specific and a generic corpus.

While the mean *idf* of the query terms in the domain specific corpus and the generic web corpus were 6.8 and 5.8 respectively, a large number of the query terms deviated significantly from the 45° line. Using the log likelihood statistic (Vogel, 2003), we measured that the deviation was significant in 86.7 percent of the terms. The data points above the 45° line, which account for about 33

percent of the query terms, represent terms that occur more frequently in the CACM corpus than a generic corpus.

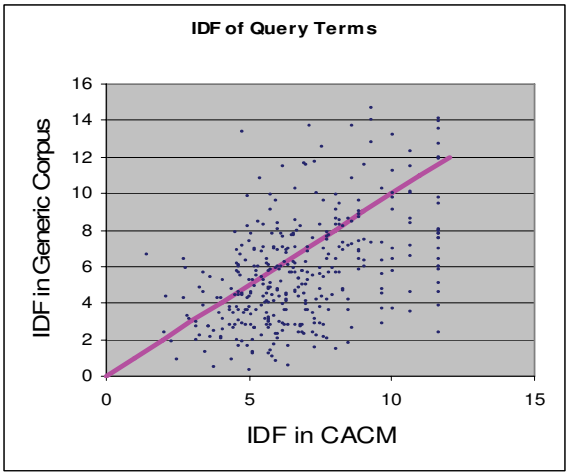

Figure 4. Scatter-plot of Term Idf in Generic and Domain Specific Corpora

We observed that most of the terms that occurred more frequently in a domain specific corpus as compared to a generic corpus were low polysemy terms, while the occurrence of high polysemy terms was much lower in the domain specific corpus as compared to their occurrence in a generic corpus. A plot of average polysemy of the terms as compared to their distance along a normal to the 45° line is given in Figure 5. This distance is a measure of the extent to which the document frequency of the term in a domain specific corpus deviates from its document frequency in a generic corpus. The distance is denoted as positive if the term lies above the 45° line and negative if it lies below the 45° line. A negative distance indicates that the term occurs less frequently in the domain specific corpus than in a generic corpus. We also observed that the terms that deviated most and were above the 45° line were domain specific terms while terms closer to the line were generic terms.

Polysemy of Domain Specific Terms

We hypothesize that query terms with high polysemy count are more likely to be generic terms while terms with low polysemy counts are more likely to be domain specific terms. Hence relevant documents are more likely to contain query terms that have low polysemy counts while non-relevant documents are more likely to be missing terms that have low polysemy counts.

We analyzed the polysemy counts of the terms occurring and not occurring in relevant and non-relevant documents. We observed that the average polysemy count of the terms not occurring in non-relevant documents was lower than the average polysemy count of the terms occurring in non-relevant documents. We analyzed the top 20 documents returned for each query and observed that over 85 percent of the retrieved documents which did not contain the query terms with the lowest polysemy counts were determined to be not relevant. Similarly, about 70 percent of the retrieved documents that were relevant contained the lowest polysemy terms from the query.

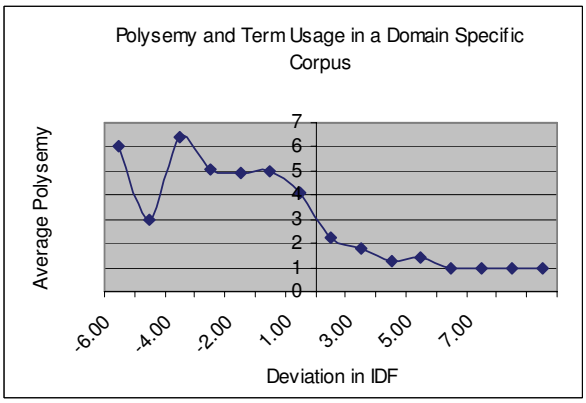

Figure 5. Average Polysemy of terms with deviation from expected idf

However, when compared with the overall percentage of relevant and non-relevant documents, the differences in the occurrence of relevant documents as a function of polysemy of the occurring and non-occurring terms were not statistically significant.

Retrieval Using Automatic Term Weighting

We developed two term weighting mechanisms, the normal distance measure and the adjusted polysemy measure, to differentiate between domain specific query terms and generic query terms. The first measure is the deviation of the inverse document frequency of the term from the expected inverse document frequency as estimated from a generic corpus. The normal distance (D) is measured as the distance along the normal between the term, denoted by the co-ordinates (idf_g, idf_c), to the diagonal given by the line idf_g=idf_c. Mathematically, the normal distance (D) is given by:

$$D = \frac{idf_g - idf_c}{\sqrt{2}}$$

Where,
idf_g is a terms inverse document frequency as estimated in a generic corpus.
idf_c is a terms inverse document frequency as estimated in a domain-specific corpus.

Apart from the above measure, we have developed a new heuristic measure, called adjusted polysemy, to determine a terms importance. The adjusted polysemy measure combines information from three different sources. We have observed that generic terms usually tend to be associated with multiple meanings and have a high

polysemy count. Therefore, the adjusted polysemy measure is inversely proportional to the polysemy of a term. In addition, we have observed that most domain specific terms have a high *idf* value in a generic corpus, as such the adjusted polysemy indicator is directly proportional to a terms *idf* in a generic corpus. However, idf_g does not measure the discriminative power of the term in the domain specific corpus. For example, although the term algorithm is a domain specific term, it occurs in over half of the documents in the CACM corpus and cannot be used to effectively discriminate between relevant and non-relevant documents. The ability of a term to differentiate between the documents of a corpus is given by the terms *idf* in the domain specific corpus. Hence the Adjusted polysemy measure is directly proportional to idf_c.

In summary, the adjusted polysemy (P_{adj}), is the product of the inverse document frequencies of the term in a domain specific and a generic corpus divided by the polysemy of the term and is given by

$$P_{adj} = \frac{idf_c \times idf_g}{Polysemy}$$

Where, idf_c and idf_g are the inverse document frequency of a term in a domain specific and a generic corpus, and *Polysemy* is the polysemy count of a term as estimated from the Wordnet thesaurus.

The precision of the best match retrieval system using the two weighting measures is shown in Figure 6. The baseline run without term weighting is labeled *MS1* and the two term weighting measures are labeled *MS2* and *MS3*. We observed that the performance of the two measures was almost identical in terms of the precision recall curves. Note that MS3 has significantly higher precision at the lower end of the recall scale. This is a significant finding because task-centric document recommendation requires very high precision in order to minimize information overload.

As a result of using the term weighting mechanism, a higher number of relevant documents were retrieved at higher threshold levels as compared to the baseline of no term weighting. We observed a 15 percent increase in average precision for the first 100 documents recommended when the term weighting mechanism was used. A comparison of the precision curves over the first 100 documents is given in Figure 7. Again, this demonstrates that the proposed query measures are useful for task-centric document recommendation because of a higher recommendation precision.

Although there is an increase in precision and recall at higher threshold levels, the precision and recall with the term weighting is much lower than the baseline run at lower threshold levels. This is because non-relevant documents that were previously retrieved at a higher threshold level are now retrieved at a lower threshold.

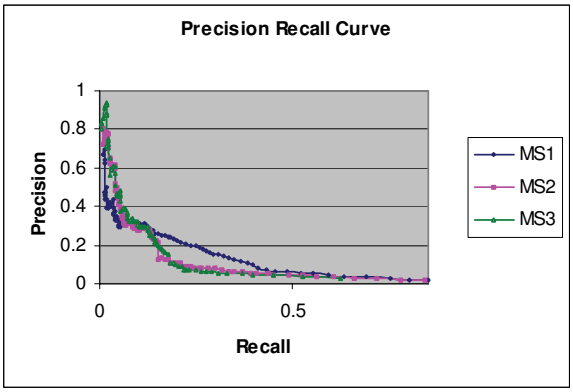

Figure 6. Precision Recall curve using term weighting measures

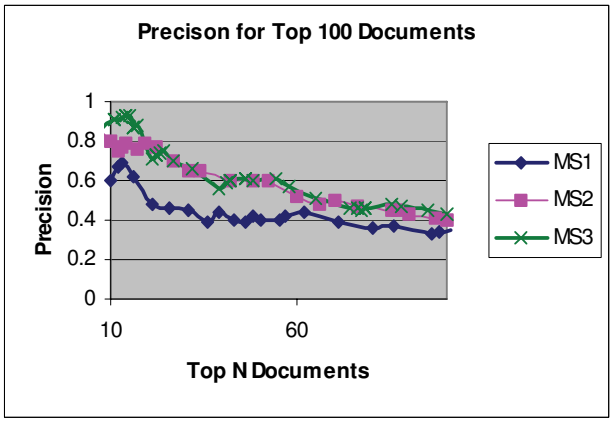

Figure 7. Precision for top 100 documents

Conclusion and Future Work

In order to prevent information overload, context-based document recommender systems need to retrieve a few but highly relevant documents. To enable the retrieval of highly relevant documents based on contextual information, we need a mechanism to differentiate between key terms and generic terms from an unstructured textual description of context.

In this paper, we proposed two measures for differentiating between key terms and generic terms in contextual information and have demonstrated the utility of the measures in a domain-specific corpus of computer science abstracts. The new term weighting measures show superior performance at higher threshold levels when compared to a standard best match retrieval system, enabling the retrieval of highly relevant documents for recommendation in context. Our approach is particularly promising for task-centric document recommendation because of a much-improved precision as compared with the general information retrieval approach.

In future work, we intend to conduct user studies to further evaluate the performance of the proposed measures. In this paper, we have evaluated the performance of the proposed measures in the CACM corpus, which is a widely used standard for evaluation of information retrieval system. The queries associated with the CACM corpus are descriptive and can be considered as being representative of contextual information. However, the use of the proposed term weighting measures in document recommender systems embedded in actual real-time systems such as email systems and workflow systems, and their effectiveness in identifying contextual information from email messages and task descriptions needs to be evaluated. In addition, we also intend to examine other several issues such as the validity of the proposed measures in multiple domain specific corpora, and the robustness of the measures when calculated using different generic corpora.

References

Abecker, A., Bernardi, A., Hinkelmann, K., Kuhn O. and Sintek, M. (2000) "Context-aware, proactive delivery of task-specific information: the KnowMore Project" *Information Systems Frontiers* **2** 3/4, pp. 253–276

Anick, P. and Tipirneni, S. (1999) "The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking" *Proceedings of the 22nd Annual International ACM SIGIR Conference*, Berkeley, CA

Apache Software Foundation (2004), Lucene 1.4.3 API, Class Similarity http://jakarta.apache.org/lucene/docs/api/org/apache/lucene/search/Similarity.html

Budzik, J. and Hammond, K. (2000) "User Interactions with Everyday Applications as Context for Just-in-Time Information Access", *Proceedings of the 2000 International Conference on Intelligent User Interfaces*, New Orleans, LA

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2002) "Placing Search in Context: The Concept Revisited", *ACM Transactions on Information Systems*, 20(1), pp116-131.

Kwan, M. and Balasubramanian, P. (2003) "KnowledgeScope: Managing Knowledge in Context", *Decision Support Systems* 35, (4), 467-486.

Krovetz, R., and Croft, B. (1992) Lexical Ambiguity and Information Retrieval, *ACM Transactions on Information Systems*, 10(2), pp115-141

Liu, Z. and Chu, W. (2005) "Knowledge-Based Query Expansion to Support Scenario Specific Retrieval of Medical Free Text" *Proceedings of the 20th Annual ACM Symposium on Applied Computing*, Santa Fe, New Mexico

Navigli, R., Velardi, P. and Gangemi, A. (2003) "Ontology Learning and Its Application to Automated Terminology Translation", *IEEE Intelligent Systems*, 18(1) pp22-31

Salton, G. and Buckley C. (1988) "Term-weighting approaches in automatic text retrieval", *Information Processing and Management*, 24(5) pp513-523

Sarnikar S. and Zhao, J. L, (2005) "A Bayesian Framework for Just-in-Time Knowledge Delivery in Web-based Business Processes", *Proceedings of the 4th Workshop on E-Business*, Las Vegas, NV, USA

Sparck Jones, K. (1972) "A statistical interpretation of term specificity and its applications in retrieval" *Journal of Documentation*, 28, pp11-21

Vogel, D (2003) "Using Generic Corpora to Learn Domain-Specific Terminology" *Workshop on Link Analysis for Detecting Complex Behavior (LinkKDD2003)*, Washington, DC