Improving UMLS Metathesaurus Query Expansion Based on the Query Specificity and Length

Xiangming Mu
School of Information Studies
University of Wisconsin-Milwaukee,
Milwaukee, WI 53211, USA
mux@uwm.edu

Kun Lu
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI 53211, USA
kunlu@uwm.edu

ABSTRACT
Query expansion is commonly used in health information retrieval. By adding synonyms and related terms using Unified Medical Language System (UMLS) Metathesaurus, query expansion can help users find more relevant information. In this paper, we studied how query specificity (the specificity of a query term) and query length (the number of words in a query term) influence the effectiveness of genomic information retrieval. Using the TREC Genomic Track 2006 test collection, our experiment demonstrated that the UMLS-based expansion on longer and more specific queries outperformed shorter and less specific queries; and the expansion using the string index method achieved better performance than using the word index method. Further case analysis revealed that for more specific queries and when using the string index method, the positive factor of adding useful related terms outweighed the negative factor of over-expanding irrelevant terms.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: query formulation

General Terms
Algorithms, Measurement, Documentation, Performance, Experimentation

Keywords
Health information retrieval; Query expansion; United Medical Language System (UMLS); Query specificity and query length; Genomic information retrieval; Biomedical information retrieval

1. INTRODUCTION
Thesaurus-based query expansion is widely used to boost the performance of information retrieval [1]. Additional synonymous, broader, related, and narrower terms are added to help retrieve more relevant documents. These relevant documents may not be indexed in the original query terms but in the expanded ones.

Research on query expansion in health and biomedical information retrieval demonstrated mixed effectiveness. Some studies have shown that query expansion was helpful [2, 3], while other research indicated that the retrieval performance was impaired because unrelated noise terms were added, reducing the retrieval precision [4-6]. It is important to further explore the factors that influence the effectiveness of query expansion in biomedical information retrieval, particularly when a large medical thesaurus such as the Unified Medical Language System (UMLS) Metathesaurus is used.

The UMLS Metathesaurus is the most comprehensive health and biomedical thesaurus containing more than 150 specific health and biomedical thesauri, including the well-known Medical Subject Headings (MeSH) thesaurus. Studies indicated that using the UMLS Metathesaurus with appropriate weighting and normalization helped improve retrieval performance [3]. However, limited research has been conducted to examine how query characteristics [6, 7] contribute to query expansion effectiveness. In this study we will explore how two query characteristics, query specificity and query length, influence the performance of four UMLS Metathesaurus-based query expansion strategies in genomic information retrieval. Details about the query expansion strategies and their performances were given in one of our previous study [6].

The remainder of the paper is organized as follows. After the literature review, we first briefly describe the definitions of the four UMLS Metathesaurus-based query expansion strategies, query specificity, and query length. In the research question section we outline our main research goals. Next in the methodology section we describe the retrieval system, dataset, tasks, and performance evaluation measurements. Findings of the study are presented in the results and analysis section. In the next discussion section we analyze and discuss the meanings and implications of our study findings. Finally we conclude the paper with a summarization and future works.

2. LITERATURE REVIEW
Automatic thesaurus-based query expansion is commonly used because it does not require a user’s conscious involvement [8, 9]. Examples of these thesauri include ProQuest Controlled Vocabulary, WordNet, MeSH, and the UMLS Metathesaurus.

After adding related and broader terms using the ProQuest Controlled Vocabulary thesaurus, Greenberg [5] reported an increased recall in her experiment. In another study using a
MEDLINE dataset, Srinivasan [10] examined three query expansion strategies using MeSH terms: expansion via an inter-field statistical thesaurus, expansion via relevance feedback, and expansion using a combined approach. The author’s study results indicated a significant 16.4% improvement in terms of 11-point average precision. As a comparative study, Aronson and Rindflesch [2] examined the same MEDLINE data with query expansion based on UMLS Metathesaurus concepts created by MetaMap. Their best results showed a 14.1% improvement over the baseline in terms of average precision. They concluded that query expansion based on the UMLS Metathesaurus was effective.

In genomic information retrieval, many studies in the TREC Genomics Track 2006 utilized various query expansion techniques to augment retrieval effectiveness [3, 11-13]. Stokes et al. [3] summarized the expansion types used by 16 TREC 2006 groups into four categories: 1) lexical variants of query terms— including plural/singular, morphological, orthographic and spelling variations; 2) term synonyms; 3) ontological relationships, including specialization/generalization associations; and 4) concurrent relationships generated from a biomedical document corpus. After applying these expansion techniques, many TREC Genomic studies reported that query expansion was helpful [3].

Not all studies concluded positive effects on thesaurus-based query expansion. Using thesaurus relationships and definitions in the UMLS Metathesaurus, Hersh et al. [4] investigated different expansion strategies by adding three types of manually-assigned Metathesaurus terms: synonymic, hierarchical, and related terms. Their results showed degraded aggregate performance in every expansion strategy. Even though no underlying explanation about the degraded performance was given by the authors, case by case analysis was suggested for further exploration. In another study, Greenberg [5] found that when related terms and broader terms were used, there was a statistically significant performance decline in terms of the retrieval precision. Recently, Mu & Lu’s study on genomic information retrieval demonstrated that the retrieval performance in terms of Mean Average Precision (MAP) declined after applying four UMLS Metathesaurus-based query expansion strategies [6].

Many studies have been conducted to explore the factors that influence the effectiveness of query expansion. Some results showed that adding appropriate terms to expand initial queries is the key to an effective query expansion strategy [14-19]. In genomic information retrieval, due to the complexity of medical terminology, it is more challenging to automatically and successfully “choose” the right terms for expansion. One proposed solution is to add weighting factors and to use normalization method [3]. Another approach is to choose the most appropriate expansion method regarding the scope and level of expansion [6]. Mu & Lu’s study found that query expansion on the string index level was more effective than that on the word index level. Unfortunately, the authors did not provide convincing evidence to fully explain such performance differences [6]. In this study, one of our research objectives is to explore this issue.

In addition to the query expansion strategy factors, which are associated with issues such as how to select an appropriate thesaurus and the extent of term expansion [3, 4, 6], the effectiveness of query expansion may also be associated with query statistic characteristics. Query statistic characteristics refer to the intrinsic nature of a query which can be, for example, query specificity, query similarity, query coherency [7], and query complexity [6].

One pioneer research investigating the impact of query characteristics was Sparck Jones’ study of query specificity based on corpus statistical analysis. Considerable improvements in search performance were observed after the query specificity was included in weighting [20]. In a later study, using WordNet synonym sets as related concept representations, Voorhees [21] illustrated that query expansion tools performed better when a query was not well described. Following this research line, Heine [22] further investigated the relative influences of three variables—database informativeness, query size and query term specificity. Her study results indicated that the number of query terms was the dominant variable influencing search performance. In total, the three variables contributed to over 50% of the search performance variations. However, the author did not provide a detailed explanation on why these factors significantly influenced the retrieval performance.

Recently, Plachouras, He, and Ounis [23] defined a query specificity related concept called Average Inverse Collection Term Frequency (AvICTF). Using the TREC Web collections (WT10G and .GOV2), the authors examined the correlation between the AvICTF and the Mean Average Precision (MAP). They found that there was a higher correlation between AvICTF and MAP for shorter queries than for longer queries. In another study using the same datasets, He & Ounis [24] proposed an adaptive query expansion strategy. When the AvICTF value was lower than a threshold, the authors classified the query as difficult and then disabled the query expansion operation. Their results indicated that the performance of this adaptive query expansion strategy depended on the relative size of the document collection [24].

Mu & Lu [6] defined the query complexity as the number of query terms. The authors studied the relationship between the query complexity and the retrieval effectiveness in genomic information retrieval. They found that simple queries significantly outperformed complex queries when UMLS Metathesaurus-based query expansion was used. The authors suggested future studies to explore the underlying reasons for such performance differences.

3. Query Specificity, Length, and Expansion Strategies

3.1 Definitions

Before we discuss particulars regarding our study, it is helpful to define the terminology we will be using, including “query,” “concept,” “term,” and “word.” In the context of this study, a query represents a user’s information need. It is taken from the user’s question and may contain multiple concepts. Each concept is represented by a query term and each term may contain multiple words. For example, topic 160 is described as “What is the role of PrnP in mad cow disease?” The query submitted to the search system is —“PrnP” and “mad cow disease”. There are two concepts in this query and each is represented by a query term. There are three words in the query term “mad cow disease” and only one word in the term “PrnP.”

3.2 Query Specificity

According to Sparck Jones, the term specificity represents “the discriminating power of individual terms” [20]. Sparck Jones argued that term specificity should be associated with its usage
frequency rather than with the semantic meaning. Even though defined in various formats, the key idea about term specificity is that the less frequently a term appears in the data corpus, the more specific the term is. This idea has been widely used for term weighting [16]. We apply the same concept to our definition of query specificity.

Suppose the data corpus has \( N \) documents. For a query with \( M \) terms and each appearing in \( q_i \) documents, the query specificity index (QSI) is defined as:

\[
QSI = \log(N + 1) - \frac{1}{M} \sum_{i=1}^{M} \log(q_i + 1)
\]

Here we use \( q_i + 1 \) to replace \( q_i \) to avoid the log0 situation. For convenience, we refer to this definition as the Query Specificity Index (QSI). This definition is compatible to other similar query specificity definitions [23] [7].

In this study, we used PubMed as the data corpus to compute QSI. At the time of our study, the total number of documents in PubMed was 18,768,606. One characteristic of this definition, however, is that it closely ties with the dataset used. So a query’s specificity varies in accordance with the data set selected.

### 3.3 Query Length

The number of words in a query term is also loosely associated with the term specificity. For example, the two-word query term “breast cancer” is more specific than the one-word query term “cancer,” a three-word term “breast cancer treatment” is even more specific. A search on “compounds” yields 512,348 results in PubMed, but a search on “organic selenium compounds” returns only 41 documents.

We define Query Length Index (QLI) as a measure of the average number of words per term in a query. Suppose a query has \( M \) terms. For the \( i \)-th term, it contains \( n_i \) words. Then QLI is defined as:

\[
QLI = \frac{1}{M} \sum_{i=1}^{M} n_i
\]

Based on this definition, the more words a query term contains, the higher the QLI score is. Similar to the definition of query complexity in a related study [6], QLI describes one statistical feature of a query. The primary merit for this definition is its simplicity and independence from the dataset used.

### 3.4 String vs. Word Index

There are two types of indices supported by the UMLS Knowledge Source Server (UMLSKS) – a string index and a word index. The string index directly connects a Metathesaurus string (e.g., “mad cow disease”) to the related CUIs (e.g., “Encephalopathy, Bovine Spongiform” [C0085209]) or LUIs (e.g., “BSE”). The word index splits the string on word boundaries and then connects each word to the related CUIs or LUIs, using Boolean “AND” [25]. For example, “mad cow disease” will connect to CUIs that are related to the term “mad” (85 unique CUIs), “cow” (238 unique CUIs) and “disease” (1345 unique CUIs). These three groups of CUIs will be combined using AND to find common CUIs as final results (examples here).

Query expansion can be implemented using either string index or word index.

### 4. RESEARCH QUESTIONS

Our primary research question is:

Do query specificity and query length influence the performance of the four UMLS Metathesaurus-based query expansion strategies in genomic information retrieval? If so, how?

And we also explore:

What are the positive and negative factors attributing to the performance differences of the four UMLS Metathesaurus-based query expansion strategies in genomic information retrieval?

### 5. RESEARCH METHODS

#### 5.1 Search System

We developed our retrieval system based on the Lemur Toolkit [26]. The standard stop list and Porter stemming were applied. For the sake of simplicity, we did not implement any further optimization techniques in the four expansion strategies.

In addition to the four UMLS Metathesaurus-based expansion strategies, we also created a baseline without query expansion as a study benchmark.

#### 5.2 Data Collection

The TREC Genomic 2006 Track data collection was used in the experiment. It contains 162,259 full-text articles derived from 49 journals through Highwire Press. A total of 24 gene related search topics were tested in the study. Each topic contains one or more biological objects, processes, and their relationships. The biological objects may be genes, proteins, or gene mutations. The biological processes include physiological processes, diseases, etc. The relationships involve the interactions between the biological objects and processes such as cases, contributions, effects, associations, and regulators [27]. An example topic is: What is the role of PrnP in mad cow disease?

#### 5.3 Search Queries

In this study we used expert-created keyword queries instead of the original natural language topic descriptions. We did this because in a real-world setting most of the search queries are composed of two to three terms instead of a natural language phrase [28]. Seven experts, including one biomedical PhD student, two nurses, two reference librarians, and the authors, examined the topics and generated the query set after a group discussion. Since all search topics from the TREC Genomics 2006 conform to the same template, we only chose the key elements (or concepts) involved to form the baseline queries and discarded other terms such as “what, role, how, contribute, etc....” For example, topic 160 states: “What is the role of PrnP in mad cow disease?” With this topic, we used the two terms “PrnP” and “mad cow disease” as the baseline query. We believe this approach is reasonable as long as the queries are consistent across different strategies in our comparison study. On average, there are 2.58 terms in the baseline query. String index expansion moderately increases the number to 5.08. Word index expansion, however, significantly increases the number to 292.

#### 5.4 Measurements

As the ground truth is provided in the data collection, we use the standard Mean Average Precision (MAP), 11-point precision/recall, and the average precision/recall at maximal recall point (the average precision/recall at the point when the last relevant document is retrieved) to evaluate the retrieval performance. A paired Wilcoxon signed-rank test at the 0.05 confidence level is used to determine statistical significance.
6. RESULTS AND ANALYSIS

In order to study the influence of query specificity and query length on the different UMLS query expansion strategies, we classified the queries into three categories. For query specificity, we have High QSI queries, Middle QSI queries, and Low QSI queries. The High QSI queries are the eight highest QSI score queries (topics 160, 164, 168, 170, 176, 177, 178, and 179); the Middle QSI queries are those with the next highest QSI values (topics 166, 169, 171, 175, 182, 183, 185, and 187); and the Low QSI queries are those with the lowest QSI values (the remaining query topics). Examples of each QSI category are given as follows:

- Topic 160 (“What is the role of PrnP in mad cow disease?”) represents an example for High QSI: “PrnP”, “mad cow disease”
- Topic 183 (“How do mutations in the NM23 gene affect tracheal development?”) represents an example for Middle QSI: “NM23 gene”, “tracheal development”
- Topic 161 (“What is the role of IDE in Alzheimer disease?”) represents an example for Low QSI: “IDE”, “Alzheimer disease”

For query length, we have High QLI queries, Middle QLI queries, and Low QLI queries. The High QLI queries are those with QLI values greater than two (query topics 161, 163, 166, 167, 169, 170, 176, 178, 179, and 187); the Middle QLI queries are those with QLI values equal to two (query topics 160, 165, 177, 182, 183, and 185); and the Low QLI queries are those with QLI values less than two and include the remaining seven query topics. Examples of each QLI category are given as follows:

- Topic 166 (“What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA?)”) represents an example for High QLI: “Transforming growth factor-beta1”, “cerebral amyloid angiopathy”
- Topic 177 (“How do Bop-Pes interactions affect cell growth”) represents an example for Middle QLI: “Bop-Pes interactions”, “cell growth”
- Topic 172 (“How does p53 affect apoptosis”) represents an example for Low QLI: “p53”, “apoptosis”

6.1 Mean Average Precision (MAP)

Our results showed that for both low and high QSI queries, MAP for the string index expansion was statistically significantly better than that for the baseline (p=0.017<0.05). We further evaluated the expansion performance by examining the average precision and recall values at the maximal recall point or, in other words, at the point when the last relevant document is retrieved. The average precision at this point reflects the retrieval performance when the retrieval exhaustivity is important. On average, we noted that improvements of recall for the High QSI queries were more than those for the Low QSI queries. We also noted that term level expansion strategies achieved higher recall than concept level expansion strategies. This might be because more terms were expanded using term level expansion strategies and thus more relevant documents were retrieved eventually.

6.2 Relationship between Query Specificity and Query Length

From the above analysis, we have noted that in many aspects the queries demonstrated similar patterns with QSI and QLI. Figure 1 illustrates the QSI and QLI values for all 24 search topics. A Pearson correlation analysis indicated that the correlation coefficient was 0.66 and it was statistically significant at the 0.01 level. In other words, higher term specificity corresponds to higher term length. As QSI is relatively more difficult to calculate and not consistent due to its dependence on the data collection, the significant positive correlation implies that we might use relatively simple QLI to approximately estimate the value of QSI.

7. DISCUSSIONS

7.1 Expansion Strategies

The positive aspect of query expansion is that it adds useful synonyms to the original queries and helps improve the retrieval recall; while the negative aspect is that it may add irrelevant or superfluous terms and thus reduce the retrieval precision. Our results showed that UMLS-based string index expansion strategies outperformed word index strategies in terms of the Mean Average Precision (MAP). One explanation is that the string index expansion frequently had more of the positive aspect of query expansion. For example, for query 181 (“Huntington gene” and “Huntington’s disease”) using SC and ST expansion strategies, we obtained a 102% and 162.8% increase in MAP as compared to the baseline. In the top 20 documents returned by the baseline, there were 15 relevant documents and 5 irrelevant ones. Using SC expansion, 18 out of the top 20 documents were relevant; using ST, all of the top 20 documents were relevant except the 17th. There were several relevant documents retrieved by SC and ST which were not found by the baseline query. A detailed check of these documents showed that they all contained the expanded term “HD Gene” which was not in the original baseline query. Therefore, in this case “HD Gene” is an effective expansion term helping retrieve relevant documents otherwise missed by the original baseline query.

Query 167 (“nucleoside diphosphate kinase” and “tumor progression”) demonstrated how over-expansion deteriorated retrieval performance. MAP of the original baseline query was 0.45. It dropped to 0.12 (72% decrease) in SC and to 0.1 (78% decrease) in ST. In the top 20 retrieved documents by the baseline query, there were 14 relevant and 6 irrelevant documents. In SC, there were 7 relevant documents and in ST there were only 6 relevant documents. Some relevant documents found by the baseline query were not found in SC or ST because other irrelevant documents were ranked higher, moving into the top 20
document list. A detailed check showed that the concept of “tumor progression” was over-emphasized after the expansion. In ST, the expanded terms for “tumor progression” such as “cancer progression” and “neoplasm progression” were found in some highly ranked irrelevant documents that did not contain the term “nucleoside diphosphate kinase.” More weight was incorrectly added to the concept relating to “tumor progression” after expansion. This made some irrelevant documents containing that concept rank higher, thus deteriorating the retrieval performance.

Comparing the four UMLS Metathesaurus-based expansion strategies, we found that using the word index approach significantly decreased the retrieval performance. Case analysis for WC and WT is more challenging since they usually expand extensively. It is more difficult to tell which expanded term leads to how much result change. A general finding from our analysis indicated that the more a query was expanded using the word index expansion, the better the chance irrelevant documents were ranked higher. We continue to use the query 181 as an example. The retrieval performance was improved after SC and ST expansion. However, when WC and WT expansion were applied, the MAP score dropped significantly from 0.49 and 0.64 to 0.005 and 0.003 respectively (98% decreases for both). One explanation might be that the word index expansion in this case added too many irrelevant terms making the negative factor dominant. In contrast, the string index expansion was less likely to have a negative effect on search results. Based on these findings we would recommend that in general we should use the string index instead of word index method for automatic UMLS Metathesaurus-based query expansion, particularly when a high recall is not the primary goal.

However, this does not mean that the word index expansion is inherently the worst strategy. The word index expansion achieved better maximal recall than the string index expansion. This means even though not ranking very well, word index expansion strategies helped find more relevant documents than string indexing strategies. This is helpful for cases when exhaustivity is essential. To improve the performance, some techniques were proposed to alleviate the negative impact of over expanding. For example, Stokes, et al. [3] indicated that using expansion word normalization techniques was helpful. Zhou et al. [13] explicitly modified the weight of expansion terms based on their strength of connection with the original query terms. Si et al. [12] allocated different weights based upon word type (aliases, synonyms, acronyms, and function word) and expansion source in their search system. These studies provided a promising starting point. Further research on how to effectively control the negative factor of adding irrelevant terms in UMLS-based expansion strategies will be interesting and helpful.

7.2 Query Length and Query Specificity

The study results showed that query characteristics interact with query expansion strategies influencing retrieval performance. Better performance was observed by applying expansion strategies on longer (High QLI) and more specific (High QSI) queries than on shorter (Low QLI) and less specific (Low QSI) queries. One explanation is that when the query specificity is high, relatively less irrelevant “noise” terms are introduced from the UMLS Metathesaurus. In addition, the added expansion terms are helpful in finding more relevant documents. For example, the baseline MAP for a High QSI query 160 (“Prnp” and “mad cow disease”) was 0.2418 before query expansion. After applying string index term level expansion (ST), the MAP score increased to 0.7564 -- which was a 212.92% improvement. In the top 20 retrieved documents by the baseline query, there were 13 relevant and 7 irrelevant documents. With ST, there were 18 relevant documents and only 2 irrelevant documents. A detailed check of those relevant documents revealed that most of the expanded terms for “mad cow disease” such as the “bovine spongiform encephalopathy” were helpful in finding relevant documents. For example, a relevant document 16141216.html, which was missed by the original baseline query, was ranked number one after the ST expansion. We believe that for the High QSI/QLI queries, the positive factor of adding useful synonyms or related terms outweighs the negative factor of introducing too many irrelevant terms. In contrast, for the Low QSI/QLI queries, the relatively more general terms tend to be over-expanded and thus have a higher opportunity to add more irrelevant terms, deteriorating the retrieval performance.

We noted that our conclusions seemed different from Voorhees’ findings that less well-developed queries achieved better performance after using query expansion [21]. However, in Voorhees’ study hand-chosen terms from WordNet were used. This is significantly different from our UMLS Metathesaurus-based automatic query expansion strategies. Using hand-chosen expansion terms guaranteed the quality of expanded terms, thus reducing the probability of adding irrelevant noise terms. Accordingly, less well-developed queries would benefit more from query expansion by adding more relevant terms. When using the automatic UMLS query expansion, less-developed or very general terms might bring in a broader range of irrelevant terms, reducing the retrieval effectiveness.

Another important issue is whether our conclusions are affected by the choice of the baseline terms. In order to test this, we recomputed our results using a second query set that was proposed by one reference librarian. Seven of the queries in the second query set differed from the one used in our study. These queries all contained the term “mutation” along with a specific gene. For example, the original query for topic 180 was “Huntingtin gene; mutation; Huntington disease.” The second query was “Huntingtin gene mutation; Huntington disease.” Obviously, the difference in the second query is that the term “mutation” was combined with the gene name. As a result, the second query set had a higher QSI and QLI. For example, the original QSI for topic 180 was 3.502. The second QSI was 4.048.

Results from the second query set drew the same conclusions as the original query set. Furthermore, we found that query expansions on the seven new queries in the second set achieved higher MAP scores. For example, after query expansion using ST strategy, the MAP score increased 1.79% using the new seven queries. This is much better than the original seven queries that led to a 45.52% decrease in the MAP score. Findings from the second query set further confirmed our conclusion that query expansion strategies work better on high QSI and QLI queries.

7.3 Additional Findings

Our results indicated that query specificity and query length were positively correlated and the correlation was statistically significant. Considering the simplicity of computing QLI and its independence on the data corpus, this finding provides us an alternative means to quickly estimate the value of query specificity. In practice, it is particularly useful for information
retrieval applications that support a dynamic data corpus. For example, the collection size of PubMed changes almost every day because of the addition of new publications. Using QL1 to approximately gauge a query’s specificity is straightforward and consistent.

7.4 Study Limitations

One limitation of the study is the experiment’s batch manner that did not include the user and user-system interaction factors. It is possible that experienced users may manipulate the search system with successful queries that any automatic expansion strategy might deteriorate. A well developed search interface supporting sophisticated user interaction functions may also imply different recommendations for the best expansion strategies. Further user-based empirical studies will help better understand these important issues. In addition, the number of search topics used in the study is relatively low. Testing our conclusion on additional data set with a broad range of search topics will be helpful in the future.

7.5 Future Studies

This study specified that the expansion of highly specific or lengthy queries led to better performance. It would be desirable if we were able to provide a threshold value indicating when we should apply query expansion. Yet, there are many other variables that also affect the expansion performance, such as differences in the search questions, thesauri, expansion methods, retrieval algorithms, and data collection. Thus, it is difficult at the current stage to establish a quantitative threshold value from the results of one study with a limited number of queries using one specific collection. In the future, more studies regarding this issue will help us to better understand the factors involved.

In this study we applied one search algorithm [26] to test the effects of query specificity and query length on different query expansion strategies. It is possible that such effects might interact with the search algorithm chosen. In the future, it will be very interesting to further examine such an issue and explore whether the conclusions continue to stand with other retrieval algorithms.

As the goal of the study was not to create a superior query expansion model, but rather to explore how query specificity and length affect four different UMLS Metathesaurus-based query expansion strategies in genomic information retrieval, we chose the simplest settings without any implementation of related query expansion optimization techniques such as structured queries [29], query normalization [3], weighted queries [30], relevance feedback [10, 31], and Bayesian machine learning [32]. In the future it will be interesting to further explore how these and other optimization techniques interact with the query characteristics to influence the biomedical information retrieval performance.

8. CONCLUSIONS

Many factors may contribute to effective query expansion in biomedical information retrieval. In this paper we investigate how query specificity and query length influenced the performance of genomic information retrieval after using four UMLS Metathesaurus-based automatic query expansion strategies. Our results from experiments on the TREC Genomic Track 2006 dataset indicated that longer and more specific queries achieved better retrieval performance than shorter and less specific queries.

The string index expansion was significantly better than the word index expansion. An in-depth case analysis revealed that adding appropriate expanding terms and avoiding over-expansion with irrelevant terms is key for successful genomic information retrieval. When queries were more specific, the positive factor of adding useful terms over-weighed the negative factor of over-expansion, resulting in better retrieval performance. Findings from our study improve our knowledge of factors influencing the effectiveness of UMLS Metathesaurus-based query expansion, and inform directions for future studies.

9. REFERENCES


