# Extending PubMed Related Article (PMRA) for Multiple Citations

Sachintha Pitigala[1] and Cen Li[2]

[1] Center for Computational Sciences, MTSU, Murfreesboro, TN, USA
[2] Department of Computer Science, MTSU, Murfreesboro, TN, USA
spp2k@mtmail.mtsu.edu, cen.li@mtsu.edu

**Abstract.** PubMed is the most comprehensive citation database in the field of biomedicine. It contains over 23 million citations from MEDLINE, life science journals and books. However, retrieving relevant information from PubMed is challenging due to its size and rapid growth. Keyword based information retrieval is not adequate in PubMed. Many tools have been developed to enhance the quality of information retrieval from PubMed. PubMed Related Article (PMRA) feature is one approach developed to help the users retrieve information efficiently. It finds highly related citations to a given citation. This study focuses on extending the PMRA feature to multiple citations in the context of personalized information retrieval. Our experimental results show that the extended PMRA feature using the words appearing in two or more citations is able to find more relevant articles than using the PMRA feature on individual PubMed citations.

**Keywords:** PubMed, Information Retrieval, Similarity Measures, PubMed Related Citations, Personalized Article Retrieval System.

## 1 Introduction

National Library of Medicine (NLM) started to index biomedical and life science journal articles in 1960's. The indexed citations were kept in the Medline citation database. Currently, NLM provides access to over 19 million citations dating back to 1946 [1]. In 1996, National Center for Biotechnology Information (NCBI) at NLM introduced PubMed citation database. PubMed provides access to over 23 million citations in the field of biomedicine [2]. Primarily, it allows free access to Medline citation database via internet. PubMed contains more citations than the Medline database covering the in-progress Medline citations, out of scope citations, "Ahead of Print" citations, and NCBI bookshelf citations. Therefore, PubMed is the most comprehensive citation database in the field of biomedicine.

Typical users of PubMed search for relevant articles to their specific research interests by entering one or more query terms on PubMeds web interface. This task has become more and more challenging due to PubMeds rapid growth of citations. Often times, too many citations were returned as a result of the query,

P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 55–69, 2014.

while many of the returned citations are not directly relevant to the information need. To improve the quality of retrieval from PubMed, NCBI and other academic and industry groups have developed many tools. Two main approaches have been used to enhance the information retrieval systems. The first approach builds supplementary tools for the original PubMed search interface. For example, PubMed advanced search feature ([3], [4]), PubMed auto query suggestions [5], PubMed automatic term mapping [6], and PubMed related article feature [7] are some of the PubMed supplementary tools to enhanced information retrieval from PubMed. The second approach builds entirely new tools which are complementary to the PubMed search interface. MedlineRanker [8], MScanner [9], PubFinder [10], Caipirini [11] and Hakia [12] are examples in this category.

Some popular approaches in building complementary tools are based on text classification methods ([8], [9], [10]), semantic based methods ([12], [13]) and special input (set of genes or set of protein names) based methods ([11], [14]). This study focuses on developing a complementary search tool to PubMed using text classification approach. Two recent PubMed complementary tools are MedlineRanker [8] and MScanner [9]. Each system starts with a set of abstracts that are known to be relevant to a query topic of interest, or information need. Then, it trains a Naive Bayes text classifier based on these abstracts. The learned text classifier is then used to find the relevant documents to the information need from the PubMed. However, in order to get good results from MedlineRanker and MScanner, at least 100 highly relevant abstracts need to be provided by the user ([8], [9]). This is a requirement that is not easily satisfiable by most users. Finding hundreds of abstracts that have been confirmed to be relevant to certain information need is a time consuming process.

It is desirable to have a document retrieval system that allows one to retrieve articles pertinent to his study, only requiring a small set of abstracts confirmed to be relevant to the information need. The ultimate goal of this study is to develop a complementary tool to PubMed, such that when given an information need, it is capable of training a text classifier using a small number of PubMed abstracts and retrieving highly relevant articles.

It is well known to the text mining community, it is difficult to train text classifier with good accuracy based on small data set. Therefore, an important step of building the proposed system is to identify a proper technique to increase the training set size, based on the small set of abstracts provided by the user. One approach to this problem is based on interactive user inputs. It asks explicit user feedback about the relevance of the articles extracted solely based on the small data set. The PubMed abstracts deemed relevant by the user are added to the training set. This process is repeated until the system gets sufficient amount of relevant articles. The search tool RefMed [15] was developed based on this multi-level relevance feedback with the RankSVM learning method. This multi-level relevance feedback method allows the user to express the user information need more thoroughly. However, this approach is also time consuming and tiresome which requires proficient knowledge about the biomedicine field. In addition, users need to be cautious about the feedback inputs. Since, less relevant abstracts admitted into the data set may decrease the accuracy of the classifier learned. After multiple iterations of inaccurate learning, the classifier may produce final results far from the initial information need.

Another approach is to increase the training set size by finding the most similar abstracts to the input seed abstracts based on document similarity. This approach is more efficient than the first approach, and it does not require user feedbacks. However, most of the standard similarity measures such as Pearson Correlation Coefficient [16], Cosine Similarity [17] are too general and not suitable for finding similar document from large databases such as PubMed. What we need is a similarity measure that can be used to find documents similar to the seed abstracts from a large database.

This paper focuses on extending the PubMed Related Article (PMRA) [7] measure for finding similar articles for multiple citations. PMRA is a well-established tool in PubMed. It is capable of finding relevant articles from the entire PubMed database for a given PubMed citation. PubMed real world log analysis shows that roughly a fifth of all non-trivial PubMed sessions used the related article feature [18]. The questions we would like to answer are:

- *"Can we find the articles similar to the set of seed abstracts by simply combining the individual PubMed related article lists of individual seed abstract?"*;
- *"Is it more accurate to combine the seed abstracts into one single super-citation and find articles that are similar to this super-citation?"*; and
- *"What is the best way to extend the PMRA method when there are multiple seed abstracts presented?"*.

In order to answer these questions we propose a number of extension approaches to PMRA, and a series of experiments that compare these approaches using the TREC 2005 genomic track data [19].

The rest of the paper is organized as the following: Section 2 discusses similarity measures used for text classification and the theory behind the original PMRA method. Section 3 presents the proposed extended PMRA methods, Dataset, Preprocessing steps and Procedure of estimating parameters. Section 4 describes the experimental results of the proposed methods on the TREC 2005 data. Section 5 draws the conclusions about the study and presents the future research directions.

## 2 Background

A similarity measure gives a formal definition to quantify the similarity between two instances. A distance measure quantifies how far apart two instances are. Some of the popular distance measures are Euclidean distance [20], City Block (Manhattan) distance [21] and Chebyshev distance [22]. Both similarity and distance measures are widely used in information retrieval, clustering algorithms and many other data mining applications.Distance/similarity measures can be divided into two broader categories as vector based and probabilistic based measures. Pearson Correlation Coefficient [16], Cosine Similarity [17], Jaccard Coefficient [23] and Tanimoto Coefficient [24] are some of the vector based similarity measures. Fidelity similarity (Bhattacharyya coefficient or Hellinger affinity) [25], PMRA method [7], bm25 ([26],[27]) are examples of probabilistic based similarity measures.

PMRA probabilistic similarity measure was used to develop the related article feature in PubMed. It finds articles similar to a chosen PubMed citation from the entire PubMed database. When a user selects a citation from the PubMed search results, the right panel of the browser window displays citations that have the highest PMRA similarity value, e.g., the closest matching, to the chosen citation. The list of the most similar citations forms the related citation list. The related citation list for each article in PubMed is pre-calculated, and pre-sorted according the PMRA value [7]. The calculation and sorting of PMRA lists are done at the back-end and PubMed is updated periodically with the new PMRA scores.

### 2.1 PMRA Method

Given that document $d$ is deemed related to one's information need, PMRA computes the relatedness of document $c$ in terms of the posterior probability $P(c|d)$, where $c$ can be any document in PubMed. Assuming a document can be decomposed into a set of mutually exclusive and exhaustive *"topics"* $s_1, s_2, ., s_N$. $P(c|d)$ can be computed as following equation 1:

$$P(c|d) = \sum_{j=1}^{N} P(c|s_j)P(s_j|d) \tag{1}$$

Expanding $P(s_j|d)$ using the Bayes theorem, we obtained equation 2.

$$P(c|d) = \frac{\sum_{j=1}^{N} P(c|s_j)P(d|s_j)P(s_j)}{\sum_{j=1}^{N} P(d|s_j)P(s_j)} \tag{2}$$

For a user selected document $d$, the denominator of equation 2 remains constant for any document $c$. Therefore, the denominator of equation 2 can be ignored and the following criteria can be used to rank documents based on their relatedness/similarity.

$$P(c|d) \propto \sum_{j=1}^{N} P(c|s_j)P(d|s_j)P(s_j) \tag{3}$$

Here $P(c|s_j)$ is the probability that the user find an interest in document $c$, given an interest in topic $s_j$. Similarly, $P(d|s_j)$ is the probability that the user find an interest in document $d$, given an interest in topic $s_j$. $P(s_j)$ is the prior probability of the topic $s_j$ i.e., the fraction of all documents that discusses the topic $s_j$. Therefore, relevance of a document $c$ to the given document $d$ can be computed by summing up the product of $P(c|s_j)$,$P(d|s_j)$ and $P(s_j)$ across all the topics [7].

In order to estimate $P(c|s_j)$,$P(d|s_j)$ and $P(s_j)$, PMRA introduced a concept called *eliteness* [7]. *Eliteness* explains whether a given document $d$ is about a particular topic $s_j$ or not. The original PMRA method assumes that each word in the PubMed citation (title, abstract and MeSH term list) represents a topic ($s_j$). Moreover, each word (term) in the PubMed citation represents an idea or concept in the document. A term $t_i$ is elite for document $d$, if it represents the topic $s_j$. Otherwise, term $t_i$ is non-elite for document $d$. Equation 4 can be derived using the *eliteness* concept and Bayes theorem [7]. Let, $E$ represent the *eliteness* of a term in document $d$, and $\bar{E}$ represent the *non-eliteness* of a term in document $d$. The probability a term is *elite* in a document is conditioned on the number of times, $k$, that term appears in the document:

$$P(E|k) = \frac{P(k|E)P(E)}{P(k|E)P(E) + P(k|\bar{E})P(\bar{E})} = \left(1 + \frac{P(k|\bar{E})P(\bar{E})}{P(k|E)P(E)}\right)^{-1} \tag{4}$$

$P(k|E)$ and $P(k|\bar{E})$ are calculated using Poisson distributions as shown in Equations 5 and 6.

$$P(k|E) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{5}$$

$$P(k|\bar{E}) = \frac{\mu^k e^{-\mu}}{k!} \tag{6}$$

where $\lambda$ is the mean of the Poisson distribution of the *elite* case for the given term, and $\mu$ is the mean of the Poisson distribution of the *non-elite* case for the given term. First, substitute equation 5 and 6 values in to equation 4. Then, applying document length normalization and algebraic manipulations to equation 4, we derived equation 7.

$$P(E|k) = \left(1 + \frac{\mu^k e^{-\mu} P(\bar{E})}{\lambda^k e^{-\lambda} P(E)}\right)^{-1} = \left[1 + \eta \left(\frac{\mu}{\lambda}\right)^k e^{-(\mu-\lambda)l}\right]^{-1} \tag{7}$$

where $l$ is the length of the document and $\eta = P(\bar{E})/P(E)$.

Then, we combine the concept of *eliteness* with the relatedness concept of two documents. $P(E|k)$ is used to estimate $P(c|s_j)$ and $P(d|s_j)$ in the $P(c|d)$ model. To efficiently calculate the similarity values, the $P(s_j)$ is estimated using the inverse document frequency of term (topic) $t_j$, $idf_{t_j}$.Then, the following weighting function and the similarity function are derived to calculate the similarity of the two documents.

$$P(c|d) \propto sim(c,d) = \sum_{j=1}^{N} [P(E|k)]_{t_j,c} \cdot [P(E|k)]_{t_j,d} \cdot idf_{t_j} \tag{8}$$

$$sim(c,d) = \sum_{j=1}^{N} \left[1 + \eta \left(\frac{\mu}{\lambda}\right)^k e^{-(\mu-\lambda)l}\right]^{-1}_{t_j,c} \cdot \sqrt{idf_{t_j}} \cdot \left[1 + \eta \left(\frac{\mu}{\lambda}\right)^k e^{-(\mu-\lambda)l}\right]^{-1}_{t_j,d} \cdot \sqrt{idf_{t_j}} \tag{9}$$

$$w_t = \left[1 + \eta \left(\frac{\mu}{\lambda}\right)^k e^{-(\mu-\lambda)l}\right]^{-1} \cdot \sqrt{idf_t} \tag{10}$$

$$sim(c,d) = \sum_{j=1}^{N} w_{t_j,c} \cdot w_{t_j,d} \tag{11}$$

where $w_t$ calculates the term weight for a given document. Similarity between the two documents is computed with an inner product of the term weights as in Equation 11.

### 2.2 Parameter Estimation in PMRA

PMRA similarity calculation requires that a number of parameters, $\lambda$, $\mu$, $\eta$, be estimated. A simplifying assumption has been made for the *elite* and *non-elite* Poisson distributions: half of the terms in the document are *elite* and the other half of the terms are *non-elite*. This assumption leads to equation 12, a model similar to the maximum entropy models used in natural language processing ([7],[28]).

$$\eta\left(\frac{\mu}{\lambda}\right) = \frac{P(\bar{E})\mu}{P(E)\lambda} = 1 \tag{12}$$

The weighting scheme expressed in the equation 10 can then be re-written as:

$$w_t = \left[1 + \left(\frac{\mu}{\lambda}\right)^{k-1} e^{-(\mu-\lambda)l}\right]^{-1} \cdot \sqrt{idf_t} \tag{13}$$

This way, PMRA reduces the number of parameters to be estimated from three to two. Medical Subject Heading (MeSH) information in Medline was used to estimate $\lambda$ and $\mu$. MeSH descriptors to each PubMed indexed citation are assigned manually by experts in the field of biomedicine. Therefore, terms in the MeSH descriptors can be considered as *elite* terms for the citations. The terms in the citation that do not appear in the MeSH descriptors are considered *non-elite* terms for the citation. The average appearance of a given *elite* term ($\lambda$) or a given *non-elite* term ($\mu$) can be calculated based on a collection of PubMed citations.

The following section explains the methodology of this study. In particularly, it explains how PMRA is extended for multiple citations, the dataset used for this experiment, and the data pre-processing steps.

## 3 Methodology

As stated in Section 1, the ultimate goal of this project is to develop an enhanced information retrieval system for PubMed that can suggest related articles based on classifier learned from small number of user-defined citations. This paper discusses our approaches to increase the training set size based on the small set of citations provided by the user. PMRA measure is extended for this purpose. We experimentally evaluate these approaches using the TREC 2005 genomic track data [19].

### 3.1 Extending the PMRA Similarity Measure

PMRA was developed to find the relevant citations for a single user selected citation. Currently, PMRA method is not directly applicable for finding the relevant citations for multiple user selected citations. Next we discuss a number of approaches to extend PMRA for multiple citations.

The straightforward way of extending PMRA for multiple citations is to combine the PubMed related article lists obtained from the individual seed citations, and sort all the derived articles according to their PMRA similarity values. We refer to this method as the Basic method. The PMRA related article list for individual citations is pre-calculated in PubMed. Therefore, the Basic method can be completed in a very short time. However, this method is not good at capturing the overall user concept or idea of the information need expressed through multiple citations.

The second approach is to combine multiple citations into a single citation and to find the relevant citations to this newly formed citation. This method is slower than the first approach because the newly formed citation is not present in the PubMed database, therefore no pre-computed list is available. But, this approach gives a better representation of the particular user information need by taking into account information present in all the user-defined citations.

There are multiple ways of combing the set of seed articles into a single citation:

- The first method, the All-inclusive method, simply combines the terms from all the seed citations;
- The second method, the Intersection method, forms the new citation by only including terms that simultaneously appeared in every single seed citation, i.e., intersection of all seed citations;
- The third method, the At-least-two method, forms the new citation by including terms appearing in at least two seed citations.

We experimentally compare the effectiveness of these four methods: the Basic method, the All-inclusive method, the Intersection method, and the At-least-two method in the Section 4. Next, the data used for the experiments is discussed.

### 3.2 TREC 2005 Dataset

A subset of TREC 2005 genomic track [19] was used in this study. In particularly, Ad-Hoc retrieval task dataset from the TREC 2005 genomic track was used. This is the same dataset used in the original PMRA experiment study [7]. It contains 50 different information needs (topics) from biologists. The entire document collection for the 50 topics contains 34,633 unique PubMed citations. Each topic corresponds to a different subset of documents ranging in size from 290 to 1356 documents. Relevance of each document to the given topic was judged by a group of scientists. According to their opinion all the documents in the document pool were labeled as: Definitely Relevant (DR), Possibly Relevant (PR) or Non Relevant (NR). The ten topics having the highest number of relevant documents (definitely relevant and possibly relevant) were used in this study. Table 1 shows the document distribution of those ten topics.

**Table 1.** Ten topics (information needs) that contain the highest number of relevant documents in the TREC 2005 genomic track dataset

| Topic ID | # Definitely Relevant documents | # Possibly Relevant documents | # Non Relevant documents | # Total documents |
|---|---|---|---|---|
| 117 | 527 | 182 | 385 | 1094 |
| 146 | 370 | 67 | 388 | 825 |
| 120 | 223 | 122 | 182 | 527 |
| 114 | 210 | 169 | 375 | 754 |
| 126 | 190 | 117 | 1013 | 1320 |
| 109 | 165 | 14 | 210 | 389 |
| 142 | 151 | 120 | 257 | 528 |
| 111 | 109 | 93 | 473 | 675 |
| 107 | 76 | 114 | 294 | 484 |
| 108 | 76 | 127 | 889 | 1092 |

### 3.3 Data Preprocessing

The TREC 2005 dataset has a list of PMID's for all the topics along with their relevance judgment to the given topic. First, all the PubMed citations for the given 50 topics were downloaded from the PubMed using the Entrenz utilities provided by the NCBI [29]. When downloaded, the PubMed citations are in the XML format. First, PubMed citation title, abstract and the MeSH terms were extracted from the XML documents. All the other information such as details about the author, affiliation data and journal information were ignored in this study. Then, the title, abstract text and the MeSH terms were tokenized into list of terms. From the citation term list, stopwords [30] and words containing only digits were removed. Next, stemming was applied to obtain a normalized term list for the citation. Finally, the normalized terms from the title and MeSH (Medical Subject Headings) terms with subheading qualifier were added again to the normalized term list to give more weight to those terms. Term list for each of the 34,633 citations was constructed using this data pre-processing procedure. These lists were used to estimate $\lambda$ and $\mu$ parameters and calculate the similarities between citations.

### 3.4 Estimating $\lambda$, $\mu$ Parameters

To estimate $\lambda$ and $\mu$ parameters, the normalized term list for a given article was divided into two sets, i.e. *elite* terms and *non-elite* terms. From the normalized term list, terms appearing only in the MeSH terms were labeled as *elite* terms for the given citation. Next, all the terms not appearing in the *elite* term list were labeled as *non-elite* terms for the given citation. This process was repeated for the entire collection of 34,633 documents to obtain the *elite* and *non-elite* term lists for each citation. An *elite* word dictionary was then created with the unique *elite* terms along with their average term frequencies. The average term frequency for a given *elite* term was calculated using equation 14.

$$atf_t = \frac{\sum_{i=1}^{N} tf_{t,d_i}}{df_t} \tag{14}$$

where, $atf_t$ is the average term frequency for the given *elite* term, $tf_{t,d_i}$ is the term frequency (number of occurrences) for the term t in the $i^{th}$ document's *elite* term list, $N$ is the total number of documents in the collection and $df_t$ is the total number of documents which has the term $t$ in it's elite term list.

The average term frequency defined in equation 14 corresponds to the Poisson mean ($\lambda$) for a given *elite* term. Similarly, the average term frequencies for the *non-elite* terms can be calculated using the *non-elite* term lists in the document collection. These *non-elite* average term frequencies corresponds to the Poisson mean ($\mu$) to the given *non-elite* terms.

### 3.5 Experiment Procedure

For our experiments, the Definitely Relevant and Possibly Relevant citations for a given topic (information need) were combined and labeled as relevant citations. Next, $n$ citations were randomly selected from the relevant citation set and labeled as user seeds. In this study, the number of user seed citations ($n$) was varied from 1 to 10 according to the experiments.

To compare the effectiveness of the four PMRA extension methods, each method is applied to derive a different related citation list based on $n$ user selected seed citations. The citations included in each resulting related citation list are then sorted in descending order based on their PMRA values. The precision of the method is computed in terms of the percentage of the top citations in the sorted list that were originally labeled as Definitely Relevant or Possibly Relevant, as shown in Equation 15.

$$Precision = \frac{\#of\, relevant\, citations}{\#of\, retrieved\, citations} \tag{15}$$

In this study, precision of the top five citations (P5), the top ten citations (P10), the top 20 citations (P20), the top 50 citations (P50) and the top 100 citations (P100) were calculated for each method. Each experiment was repeated ten times by randomly selecting different seed citations from the relevant set. The following section presents the experiment results obtained.

## 4 Results and Discussion

We experimentally compare the effectiveness of the four methods: the Basic method, the All-inclusive method, the Intersection method, and the At-least-two method, in finding related citations for a given set of seed citations. Ten experiments were conducted for each information need (topic) by changing the initial seed set size $n$ from 1 to 10. Each experiment was repeated 10 times with different random seeds. Table 2 compares the overall average P5 precision of the four methods for each of the 10 information needs.

**Table 2.** Average P5 precision of the four methods for the ten information needs. The average was calculated over ten different seed set sizes. The boldface values indicate the highest P5 result for the given information need.

| Topic ID | Basic method | All-inclusive method | Intersection method | At-least-two method |
|---|---|---|---|---|
| 117 | 0.59 | 0.56 | **0.74** | 0.73 |
| 146 | 0.71 | 0.75 | **0.86** | **0.86** |
| 120 | 0.66 | 0.64 | **0.93** | 0.87 |
| 114 | 0.48 | 0.38 | 0.60 | **0.67** |
| 126 | **0.27** | 0.16 | **0.27** | 0.26 |
| 109 | **0.95** | 0.87 | 0.89 | 0.94 |
| 142 | **0.77** | 0.68 | 0.63 | 0.71 |
| 111 | **0.59** | 0.57 | 0.53 | 0.58 |
| 107 | 0.67 | 0.58 | 0.25 | **0.68** |
| 108 | 0.28 | 0.25 | 0.28 | **0.48** |

Results in Table 2 show that the Basic method produced comparable results to the results from the original PMRA study [7]. This makes the Basic method a good baseline to compare the other three methods. Results from this experiment also show that overall, the At-least-two method produced more relevant documents within the first five documents in the final output. For the 10 information needs, it has the highest P5 value, or when it isn't the highest, its P5 values are very close to the highest value. This is because the At-least-two method captures important words from different seeds and produce more specific combined citation for given information need.

The results also show that the All-inclusive method is not an effective method for combining multiple citations. It forms the new citation by taking into it all the words from all the seed citations. This causes the length, $l$, of the new citation to become much greater than that of a regular citation. Higher $l$ value in equation 13 reduces the weighted values of the words in the combined citation. Therefore, this method produced less accurate results compared to other three methods.

Next, precision of the top 50 citations (P50) and precision of the top 100 citations (P100) were used test the effectiveness of four methods. Tables 3 and 4 present the average P50 and P100 precision for the 10 information needs respectively.

Tables 3 and 4 show the Intersection method and At-least-two method outperform the Basic and the All-inclusive methods. In fact, the Intersection method and At-least two method were able to produce related citation lists that are 20% more accurate than those produced by the other two methods for a number of topics. Also, the All-inclusive method produced the worst results using P50 and P100 precision measures.

**Table 3.** Average P50 measure of ten information needs. The average was calculated using ten different seed set sizes. The boldface values show the highest P50 result for the given information need.

| Topic ID | Basic method | All-inclusive method | Intersection method | At-least-two method |
|---|---|---|---|---|
| 117 | 0.55 | 0.49 | 0.68 | **0.72** |
| 146 | 0.58 | 0.56 | **0.76** | 0.75 |
| 120 | 0.60 | 0.51 | **0.84** | 0.78 |
| 114 | 0.38 | 0.31 | 0.49 | **0.55** |
| 126 | 0.21 | 0.19 | **0.28** | 0.26 |
| 109 | 0.83 | 0.71 | 0.78 | **0.87** |
| 142 | **0.59** | 0.53 | 0.57 | 0.54 |
| 111 | 0.45 | 0.41 | 0.40 | **0.46** |
| 107 | 0.50 | 0.44 | 0.50 | **0.53** |
| 108 | 0.36 | 0.30 | 0.21 | **0.41** |

**Table 4.** Average P100 measure of ten information needs. The average was calculated using ten different seed set sizes. The boldface values show the highest P100 result for the given information need.

| Topic ID | Basic method | All-inclusive method | Intersection method | At-least-two method |
|---|---|---|---|---|
| 117 | 0.53 | 0.47 | **0.72** | 0.70 |
| 146 | 0.51 | 0.48 | **0.74** | 0.69 |
| 120 | 0.55 | 0.45 | **0.72** | **0.72** |
| 114 | 0.34 | 0.28 | 0.45 | **0.50** |
| 126 | 0.20 | 0.17 | **0.29** | 0.25 |
| 109 | 0.74 | 0.58 | 0.74 | **0.78** |
| 142 | 0.48 | 0.42 | **0.54** | 0.46 |
| 111 | 0.40 | 0.35 | 0.34 | **0.42** |
| 107 | 0.42 | 0.37 | **0.49** | 0.45 |
| 108 | 0.31 | 0.25 | 0.17 | **0.33** |

From the results shown in Tables 3 and 4, the Intersection method and At-least-two method appear comparable across the 10 information needs. To determine which of these two methods is better, a statistical analysis is performed. The overall average and the 95% confidence interval for each method were calculated using all the experiments conducted in this study. 10 experiments were performed for each information need by changing seed set size from 1 to 10. For each seed set size, average of P5, P10, P20, P50, and P100 values were recorded. Therefore, 50 average measurements were calculated for each information need. Table 5 shows the average and the confidence interval for each of the four methods.

**Table 5.** The overall average and its 95% confidence interval of each method in this study

| | All-inclusive method | Basic method | Intersection method | At-least-two method |
|---|---|---|---|---|
| Average | 0.48 | 0.54 | 0.57 | 0.62 |
| 95% CI | (0.463, 0.497) | (0.523, 0.557) | (0.549, 0.591) | (0.602, 0.638) |

Results in Table 5 clearly show that the At-least-two method produced the best results in this study. Its 95% confidence interval is higher than that from the other three methods. The All-inclusive method performed poorly in this study. The average accuracy of the Basic method and the Intersection method are quite close. The confidence intervals of these two methods overlapped. This indicates that the performance of the Basic method and the Intersection method are mostly similar.

To take a closer look at the performance of the At-least-two method, the average accuracies are plotted for different information needs across different initial seed set sizes. In Figure 1 and 2, accuracy was calculated using the average P5, P10, P20, P50 and P100 values for a given seed set size. Figure 1 shows the accuracy change over the initial seed set size for the first five information needs, and Figure 2 shows the accuracies over different seed set sizes for the next five information needs. It is observed that, for the majority of the information needs, the optimal performance is achieved when the seed set size is between 2 and 4. Afterwards, the accuracy values level off.

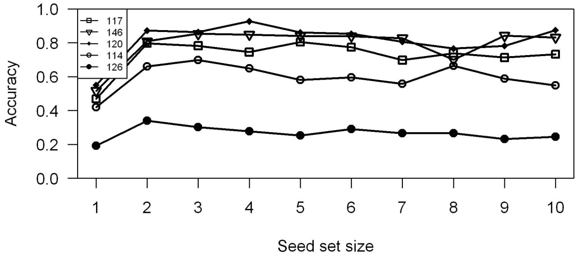

**Fig. 1.** Average accuracies for the first five information needs (117, 146, 120, 114, and 126) over different seed set sizes. Each data point is calculated using the P5, P10, P20, P50 and P100 measures for the given seed set size, and each P measure was calculated using 10 different random experiments.

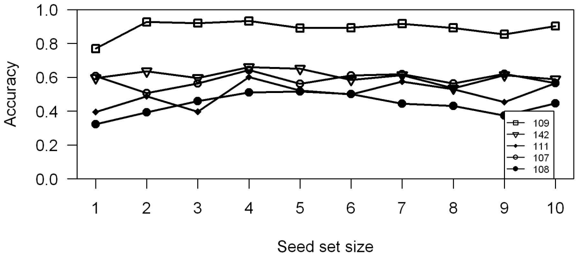

**Fig. 2.** Average accuracies for the next five information needs (109, 142, 111, 107 and 108) over different seed set sizes. Each data point is calculated using the similar procedure in figure 1.

## 5 Conclusions

This paper discusses the need to extend the PMRA similarity measure to work with multiple citations, and presents four ways to extend the original PMRA measure. To study the effectiveness of these methods, comparative analysis was conducted using the ten information needs in the TREC 2005 genomic track dataset [19].

Accuracy of the first five documents (P5) in the Basic method was comparable to the results given in the original PMRA study [7]. The At-least-two method is the best method to extend PMRA for multiple citations. This method best captures the important terms and discards the less important terms from the seed documents. In contrast, the Intersection method uses the terms appear in all seed documents, leading to a much smaller combined citation. The small number of terms from the combined citation is insufficient in accurately capturing similarity between citations. The performance of this method is comparable to that of the Basic method in this study. Overall, the All-inclusive method produced the least accurate results. The combined citation from this method is very long and contains a lot of less frequent terms. These less frequent terms and the length of the citation contributed to the low accuracy in finding the related citations.

The At-least two method generally achieved its maximum accuracy with only 4 seed citations. Adding more seed citations doesn't help to increase the accuracy under this experiment setting. The seed documents for a given experiment were randomly selected from the definitely relevant and possibly relevant documents. Therefore, a possibly relevant document which is less relevant to the information need can be selected to the seed set. This document will diverge the information need. Because of this, final relevant citation list can be less accurate. But, in practical situations, user can avoid this by selecting only relevant articles to the seed set. Future study can be conducted to reduce noise in the seed documents.

## References

1. Fact Sheet-Medline, U.S. National Library of Medicine: http://www.nlm.nih.gov/pubs/factsheets/medline.html (retrieved August 25, 2013)
2. Pubmed, U.S. National Library of Medicine: http://www.ncbi.nlm.nih.gov/pubmed (retrieved September 20, 2013)
3. PubMed Advanced Search Builder, http://www.ncbi.nlm.nih.gov/pubmed/advanced
4. Chapma, D.: Advanced Search Features of PubMed. J. Can. Acad. Child Adolesc. Psychiatry 18(1), 58–59 (2009)
5. Lu, Z., Wilbur, W.J., McEntyre, J.R., Iskhakov, A., Szilagyi, L.: Finding Query Suggestions for PubMed. In: AMIA Annu. Symp. Proc. 2009, pp. 396–400 (2009) (published online November 14, 2009)
6. PubMed's Automatic Term Mapping Enhanced. NLM Tech Bulletin (341), e7 (November-December 2004)
7. Lin, J., Wilbur, W.: Pubmed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics 8, 423 (2007)
8. Fontaine, J.F., Barbosa-Silva, A., Schaefer, M., et al.: MedlineRanker: flexible ranking of bio-medical literature. Nucleic Acids Res. 37, W141–W146 (2009)
9. Poulter, G., Rubin, D., et al.: MScanner: a classifier for retrieving Medline citations. BMC Bioinformatics 9, 108 (2008)
10. Goetz, T., Von Der Lieth, C.-W.: PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. Nucleic Acids Res. 33, W774 (2005)
11. Caipirini: using gene sets to rank literature. BioData Min. 5(1), 1 (2012)
12. Hakia (2008), http://pubmed.hakia.com/ (date last accessed September 28, 2013)
13. Wang, J., Cetindil, I., Ji, S., et al.: Interactive and fuzzy search: a dynamic way to explore MEDLINE. Bioinformatics
14. Quertle (2009), http://www.quertle.info (date last accessed September 28, 2013)
15. Yu, H., Kim, T., Oh, J., et al.: Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. BMC Bioinformatics (2010)
16. Pearson, K.: Contributions to the mathematical theory of evolution, III, Regression, heredity, and panmixia. Philos. Trans. R. Soc. Lond. Ser. A 187, 253–318 (1896)
17. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, pp. 120–122. Cambridge University Press (2008)
18. Lin, J., DiCuccio, M., Grigoryan, V., Wilbur, W.J.: Exploring the Effectiveness of Related Article Search in PubMed. Tech. Rep. LAMP-TR-145/CS-TR-4877/UMIACS-TR-2007-36/HCIL-2007-10. University of Maryland, College Park, Maryland (2007)
19. Hersh, W.R., Cohen, A.M., et al.: The Fourteenth Text Retrieval Conference (TREC 2005) NIST. TREC 2005 Genomics track overview (2005)
20. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn., p. 187. Wiley-Interscience, New York (2001)
21. Krause, E.F.: Taxicab Geometry: An Adventure in Non-Euclidean Geometry
22. Cantrell, C.D.: Modern Mathematical Methods for Physicists and Engineers. Cambridge University Press (2000)
23. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining (2005)
24. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn., p. 602, 605, 606. Academic Press, New York (2009)

25. Cha, S.: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of Mathematical Models and Methods in Applied Sciences 1(4), 300–307 (2007)
26. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the 3rd Text REtrieval Conference, TREC-3 (1994)
27. Sparck Jones, K., Walker, S., Robertson, S.E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (Parts 1 and 2). Information Processing and Management 36(6), 779–840 (2000)
28. Berger, A., Pietra, S.D., Pietra, V.D.: A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics 22, 39–71 (1996)
29. NCBI. Entrez Programming Utilities Help (2010), http://www.ncbi.nlm.nih.gov/books/NBK25501/
30. PubMed Stopwords, http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_170.html (date last accessed August 24, 2013)