Evaluating Different Ranking Functions for Context-Based Literature Search

Nattakarn Ratprasartporn, Sulieman Bani-Ahmad, Ali Cakmak, Jonathan Po, Gultekin Ozsoyoglu
Department of Electrical Engineering and Computer Science
Case Western Reserve University, Cleveland, Ohio 44106
{nattakarn, sulieman, cakmak, jlp25, tekin}@case.edu

Abstract

Context-based literature digital library search is a new search paradigm that creates an effective ranking of query outputs by controlling query output topic diversity. We define contexts as pre-specified ontology-based terms and locate the paper set of a context based on semantic properties of the context (ontology) term. In order to provide a comparative assessment of papers in a context and effectively rank papers returned as search outputs, prestige scores are attached to all papers with respect to their assigned contexts. In this paper, we present three different prestige score (ranking) functions for the context-based environment, namely, citation-based, text-based, and pattern-based score functions. Using biomedical publications as the test case and Gene Ontology as the context hierarchy, we have evaluated the proposed ranking functions in terms of their accuracy and separability. We have found that text-based and pattern-based score functions yield better accuracy and separability than citation-based score functions.

1. Introduction

At the present time, search queries in literature digital libraries either lack or do not provide effective paper-scoring/ ranking functions. We argue that the main reason for the ineffectiveness of ranking functions is that they do not take into account the diversity of papers returned as the output of keyword-based search queries. Without an effective scoring and ranking system, users are forced to scan a large paper set and potentially miss important papers. As an example, PubMed [1], which contains more than 14 million publications, does not have a paper-scoring/ranking system. Instead, PubMed simply lists search results in descending order of their PubMed ids or publication years. Other well-known digital libraries, such as ACM Portal [26] and Google Scholar [27], use only simple text-based and/or citation-based scores to rank search results.

In an earlier work [2], we proposed a new literature digital library search paradigm, *context-based search*, that controls the diversity of search input topics and effectively ranks query output publications. Before query submission, two query independent pre-processing steps are performed: assign publications into pre-specified ontology-based contexts; and compute *prestige (importance/ranking) scores* for papers with respect to their assigned contexts. Thus, in a given context, a paper with high prestige score is highly relevant to the context. Then, at search time, (a) only those papers in contexts of interest are involved in the search, and (b) search results in each context are ranked by their *relevancy scores*. A paper's relevancy score in a context is a combination of the paper's pre-computed prestige score (based on the context) and the paper-to-query matching score. Contrasting other search paradigms, context-based search output is enhanced by a context-based paper classification, which eliminates the problem of topic diffusion and reduces output size [2]. Since only semantically related papers in contexts of interest (as opposed to all papers) are involved in the search, search output ranking is more consistent and accurate. In [2], we tested our search paradigm by using PubMed [1] papers as a testbed and Gene Ontology (GO) [3] as a context hierarchy. When compared with PubMed keyword-based search engine query results, the context-based search approach was shown experimentally [2] to reduce the query output size by up to 70% and increase the search result accuracy by up to 50%.

As described above, implementing the context-based search involves five tasks: (1) assign papers to contexts, (2) compute prestige scores for papers in each context, (3) locate search contexts for a given keyword-based query, (4) perform search, and (5) rank search results. Tasks 1, 3, 4, and 5 have been extensively studied in our previous work [2]. This paper investigates task 2 as follows:

- We present three different context-based prestige score functions, namely, *citation-based*, *text-based*, and *pattern-based* score functions. As mentioned above, to rank search results within a given context, we use (a) prestige scores of papers in the context, and (b) similarity scores between the search query and the papers. The citation-based function employs the well-known PageRank algorithm [8-10], which recursively determines the prestige of a paper using citations to the paper and scores of papers citing the paper. While the citation-based score function uses only citation information, the text-based score function utilizes paper's content, authors, and citations as follows. First, a paper that best characterizes the context is selected as a *representative paper* of the context. Then, the text-based prestige score of a paper *p* in context c_i is computed from (a) text-based content similarity, (b) author overlap, and (c) citation similarity between

1-4244-0832-6/07/$20.00 ©2007 IEEE.

p and the representative paper of c_i. Compared with the other two scoring functions, the newly proposed pattern-based score function constructs patterns based on the context's *identifying elements*: Prestige score of a paper is assigned using two criteria, namely, (a) the confidence that a pattern represents the context, and (b) the matching strength between the paper and the pattern.

- We use *accuracy* and *separability* to evaluate the three prestige score functions. Score function accuracy relies on the precision scores resulting from the keyword-based search. Also, assuming that multiple score functions agreeing in their top-k paper scores are accurate [11], we measure the *top-k overlapping ratio* between each pair of score functions. Separability, which refers to the score distribution in a context, is desired to be uniformly distributed for prestige scores within a context.

For empirical evaluation, approximately 72,000 full-text PubMed papers from the genomics area were completely parsed and assigned to one or more Gene Ontology terms (contexts). For small-sized contexts (i.e., ≤ 100 papers), paper prestige scores are potentially misleading and were excluded from the experimental results.

We summarize our experimental findings as follows:

- *In the context-based environment, text-based and pattern-based scores yield better accuracy than the citation-based scores.*

This finding is not consistent with the web-based (non-context-based) environment where citation-based scores are known to be more accurate. One possible explanation is that papers of some contexts cite or are cited by large numbers of papers outside the contexts. This causes the citation graphs to be sparse within those contexts, which negatively affects the accuracy of the citation-based score function. Another possible reason is that citing and cited papers may not be topically related to each other. Therefore, citations in a context may not always indicate that citing/cited papers are important with respect to the context. We also observe that, as we drill down in the context hierarchy, context size and diversity decrease. As a result, the citation-based score accuracy is reduced due to small-sized contexts, while the text-based and the pattern-based score accuracy improves.

- *For score separability in the context-based environment, text-based, pattern-based, and citation-based functions are ranked from best to worst.*

As we drill down in the context hierarchy, the text-based score separability improves, while the pattern-based and the citation-based score separability are reduced. However, in high-level contexts, pattern-based scores possess better separability than citation-based scores.

Section 2 describes desirable properties of a prestige score function. Section 3 explains the three prestige score functions. Sections 4 and 5 present the experimental setup and experimental results. Section 6 summarizes the related work. Section 7 concludes.

2. Desirable Features of Prestige Score Functions

We use *accuracy* and *separability* [11] to evaluate the quality of a score function. Accuracy can be measured using recall and precision scores of given queries. When searching the web, most web search engine users stop looking at search results after the second page [20]. Similarly, it is not feasible for users to scan a large number of search results for a large literature digital library domain (e.g., PubMed). Therefore, high recall for a large number of returned papers is less significant than high precision for high-ranking papers. Hence, we use only precision to evaluate the accuracy of score functions. The precision is defined as:

$$Precision_t = \frac{|S_t \cap R_t|}{|S_t|}$$

, where S_t is the result set of the search for query term t, and R_t is the true answer set.

While R_t may be manually decided by experts for some queries, such an approach precludes using large numbers of queries. To automatically determine R_t without any expert help, we found the A(rtificially)C(onstructed)-answer set of a query [2]. The AC-answer set of a given query is located as follows. First, a standard keyword-based search with a high threshold is used to find an initial answer set. The initial set is enlarged using text-based and citation-based expansion. In the text-based expansion, papers that are sufficiently similar to the centroid of the initial paper set are added to the AC-answer set. For the citation-based expansion, papers in the citation path of length at most 2 from the initial paper set and with high citation scores are included in the AC-answer set. We consider only the paths of length up to 2 because longer paths usually lose context and become less relevant. To ensure the accuracy of the AC-answer set, we manually verified its correctness for some sample queries [2].

An alternative approach to measuring accuracy involves comparing score function output. Assuming that multiple score functions indicate a paper as important [11], score functions agreeing in their top k paper score rankings are consistent with each other and may be considered "accurate". Top-k overlapping ratio between two functions S_1 and S_2 in context C_i is defined as follows:

$$TopKOverlappingRatio(S_1,S_2) = \frac{|P_{S1\text{-}TopK} \cap P_{S2\text{-}TopK}|}{K}$$

where $P_{Sj\text{-}TopK}$ is the set of papers with k highest S_j scores in C_i. $P_{Sj\text{-}TopK}$ may include more than k papers if a set of papers P_m has the same S_j score as the kth paper's score. In this case, we include P_m in $P_{Sj\text{-}TopK}$ paper set and change the denominator K to $min(|P_{S1\text{-}TopK}|, |P_{S2\text{-}TopK}|)$.

Since only three score functions are involved in the evaluation, we do not use the top k overlapping ratio to find the majority agreement of the score functions. Instead, we use the overlapping ratio to measure the changes in accuracy as we drill down in the context hierarchy. In other words, only the overlapping ratio changes for the same pair of score functions at different context levels are used to determine whether the score accuracy decreases or increases as the context levels become higher.

Separability of a score function refers to the score distribution within a context. To provide a comparative assessment of papers in a context, the paper scores in the context should be evenly spread (i.e., uniformly distributed). If a score function produces only a few unique paper scores for a context, a large number of papers in the context receive the same (or a very similar) score. Since papers with the same scores are considered equally important, this negatively affects separability and the ranking of search results.

3. Prestige Score Functions

This section describes three score functions, namely, citation-, text-, and pattern-based functions, used to assign prestige scores to papers in each context. In order to rank papers returned from the context-based search, we compute a *relevancy score* for each paper in the query result. The relevancy score of paper p to query q in context c_i, $R(p, q, c_i)$, is computed as the combination of the precomputed *prestige score* of p with respect to c_i, and the *text-matching score* between p and q. $R(p, q, c_i)$ is defined as [2]:

$$R(p,q,c_i) = w_{prestige} \cdot Prestige\_Score(p,c_i) + w_{matching} \cdot Text\_Matching\_Score(p,q)$$

where $prestige\_score(p, c_i)$ is the prestige of p with respect to c_i as defined by the score functions in this section, $text\_matching\_score(p, q)$ computes the similarity between p and q, and $w_{prestige}$ and $w_{matching}$ are weights of the prestige score and the text matching score.

Since contexts are represented hierarchically, a paper p can reside in both context c_i and c_i's descendant contexts. Compared to c_i, c_i's descendant contexts are more specific, and the descendant contexts' paper sets are less diverse. Hence, a high prestige score for p in c_i's descendant contexts means that p is highly relevant to c_i. Keeping this in mind, p's score in context c_i is modified to max(s_j), j ∈ {i, k, …, n} when p resides in (a) context c_i with score s_i; and (b) descendant contexts $c_k...c_n$ of c_i with scores $s_k,...,s_n$.

3.1. Citation-Based Prestige Score Function

PageRank and Hyperlink-Induced Topic Search (HITS) algorithms can be used in paper score computation [8-10]. For a web page to be "prestigious" in terms of PageRank, other "prestigious" web pages must hyperlink to that page. By substituting a paper for a web page, a paper p's PageRank score is recursively determined by the number of citations to p and the scores of papers citing p. HITS is based on "authorities" and "hubs". A paper's authority score is proportional to the total agglomerative score of hubs that cite the paper. A paper's hub score is proportional to the total agglomerative score of authorities that are cited by the paper. Previous experiments on the ACM SIGMOD Anthology [11] showed that HITS and PageRank scores are highly correlated.

We chose to implement a variation of the PageRank scoring algorithm for our experiments. Assuming citation relationships between papers in different contexts would erroneously boost citation-based paper scores with respect to a context, only citation information between papers in the given context is used for the prestige score computation. As an example, assume (i) a paper p resides in contexts c_1 and c_2, (ii) large numbers of papers in c_1 cite p, and (iii) only a few papers in c_2 cite p. Based on these assumptions, p should be considered more important in c_1 than in c_2, and citations in c_1 should not be involved in the score computation for c_2.

For each context, a paper's PageRank score is computed recursively as:

$$P_{i+1} = (1 - d)M^T P_i + E$$

where P_i and P_{i+1} are the current and next iteration PageRank vectors. The citation matrix C is an N x N adjacency matrix of a graph with papers in the given context representing nodes and citation relationships representing edges. N is the number of papers in the context. M is derived from C by normalizing all row-sums in C to 1. d is the probability that a person reading a paper p_1 will next read a paper cited by p_1, and (1-d) is the probability that he/she will next read a random paper. To guarantee PageRank algorithm convergence, a hidden citation link between a paper and all other papers E is assumed to exist. One choice is $E_1 = d$. Another option is $E_2 = (d/N)[1_N]P_i$, where 1_N is N ones vector.

3.2. Text-Based Prestige Score Function

The text-based prestige score of paper p in context c is computed using text-based similarity measures based on the Term Frequency-Inverse Document Frequency (TF-IDF) model [6] between c and p. However, contexts are represented as short phrases (e.g., GO term), which are much shorter than papers. We use representative papers of contexts instead of context terms to represent contexts. In each context c_i, papers in c_i that are highly similar to the representative paper of c_i receive high prestige scores. The text-based paper score is defined as:

$$Sim(P_X, P_C) = \sum_i weight_i * Sim_i(P_X, P_C),$$

where P_C is the representative paper of context C, P_X is a paper in C, i ∈ {title, abstract, body, index term, authors, references}, and $weight_i$ is the corresponding similarity weight.

Sim_{title}, $Sim_{abstract}$, Sim_{body}, and $Sim_{index\ terms}$ are computed using cosine similarity of the TF-IDF model [6]. Author similarity between two papers ($Sim_{authors}$) relies on two notions: *Level-0-Author-Overlap*, which occurs when two papers share common authors; and *Level-1-Author-Overlap*, which occurs when each paper's authors co-write a third paper. Paper author similarity ($Sim_{Authors}$) is defined as follows [7]:

$$Sim_{Authors}(P_Q,P_X) = L0Weight*Sim_{Level\text{-}0\text{-}Author}(P_Q,P_X) + L1Weight*Sim_{Level\text{-}1\text{-}Author}(P_Q,P_X)$$

where P_Q and P_X are papers, $Sim_{Level\text{-}i\text{-}Author}$ is the Level-i-Author-Overlap score, and *LiWeight* is the Level-i-Author-Overlap weight with $0 \leq \text{LiWeight} \leq 1$. $i \in \{0, 1\}$.

Citation similarity between two papers relies on *bibliographic coupling* and *co-citation* [7]. Bibliographic coupling [15] gives higher similarity scores to papers with common citations. Co-citation [14] gives higher similarity scores to papers that are cited by the same paper. Citation similarity ($Sim_{References}$) is defined as follows [7]:

$$Sim_{References}(P_Q,P_X)=BibWeight*Sim_{bib}(P_Q,P_X)+ (1\text{-}\ BibWeight)*Sim_{coc}(P_Q,P_X)$$

where P_Q and P_X are papers, Sim_{bib} and *BibWeight* is the bibliographic coupling score and weight, Sim_{coc} and 1-*BibWeight* is the co-citation score and weight.

3.3. Pattern-Based Prestige Score Function

Patterns of each context are constructed from the training paper set (i.e., GO annotation evidence papers for GO-specific contexts) of that context [2]. Words/terms related to a context term are considered *significant terms (phrases)* for that context. For the pattern construction phase, significant terms are constructed from two sources: (i) words in the context term, and, (ii) *frequent terms (phrases)* in the training papers. During the frequent phrase construction, significant terms from each source are combined using a procedure similar to the apriori algorithm [5].

A (regular) pattern consists of three tuples [4]: <LEFT><MIDDLE><RIGHT> where each tuple is a set of words. <MIDDLE> tuple contains only a sequence of significant term words. <LEFT> and <RIGHT> are the word sets surrounding the significant term words. By virtually walking from one pattern to another, two types of extended patterns are constructed [4]: (i) side-joined and (ii) middle-joined patterns. A side-joined pattern is created from an overlap between the left tuple of one pattern and the right tuple of another pattern. E.g., if P1 = <A><B><C> and P2 = <C><D><E>, then the side-joined pattern P3 = <A><B><C><D><E> is constructed. A middle-joined pattern is created when there is an overlap between the middle tuple of one pattern and the left/right tuple of another pattern. E.g., if P1 = <A><B><C> and P2 = <D><E><F>, and {<C> ∩ <E>} ≠ ∅, then the middle-joined pattern P3 is <A><B>{<C> ∩ <E>}<F>.

We assign pattern-based prestige scores using the following criteria: (1) confidence that a pattern represents the context, and (2) matching strength between a paper and a pattern. The pattern-based prestige score is defined as:

$$Score(P) = \Sigma_{pt\in Ptr(P)}\, Score(pt)*M(P,\ pt)$$

where *Ptr(P)* is the set of patterns that match to paper *P*, *Score(pt)* is the score of pattern *pt*, and *M(P, pt)* is the matching strength of pattern *pt* in paper *P*.

M(P, pt) is influenced by the (1) paper section containing the pattern match and (2) similarity between the pattern (i.e., *pt*) and the matching phrase in *P*.

Based on the pattern type of *pt, Score(pt)*, is computed as follows:

- *pt* is a *regular pattern*: We compute the regular pattern score based on the following middle tuple properties: (1) [*MiddleTypeScore*]: Middle tuples, which can consist of only frequent terms, only words in context term, or both frequent and context terms, receive the high, higher, and the highest scores, respectively. (2) [*TotalTermScore*]: Context term words with higher *selectivity* receive a higher score. Selectivity describes the word's occurrence frequency among all context terms. (3) [*PaperCoverage*]: A pattern's score is inversely proportional to the middle tuple frequency among all the database papers. (4) [*PatternPaperFreq*]: Higher scores are assigned to patterns whose middle tuples are frequent in the context term's training papers. The regular pattern score function is defined as follows:

$$RegularPatternScore = BaseScore * (1/PaperCoverage)^t$$

$$BaseScore = MiddleTypeScore+TotalTermScore + c*(PatternOccFreq+PatternPaperFreq)$$

where *t* and *c* are constants.

- *pt* is an *extended pattern*: Side-joined patterns score is defined as:

$$Score(Side\text{-}joined\ Pattern) = (\ Score(Pattern1) + Score(Pattern2)\ )^2$$

where pattern 1's right tuple overlaps pattern 2's left tuple. Middle-joined Pattern Score is defined as:

$$Score(Middle\text{-}joined\ Pattern) = DOO1*Score(Pattern1)+DOO2*Score(Pattern2)$$

The *DegreeOfOverlap* (*DOO*) represents the proportion of a pattern's middle tuple included in the other pattern's left/right tuple (see [4] for more details).

4. Experimental Setup

We downloaded, parsed, and populated our database with 72,027 full-text PubMed papers. All selected papers came from the genomics area, which constitutes a "semantically related" subset of PubMed papers related to GO [12].

Our experiments utilize two context paper sets generated in an earlier work [2]:

- *Text-based Context Paper Set* was created by using the text-based similarity measure between a representative paper of a context and papers in our database. Text- and citation-based scores were assigned to papers in all contexts. We did not evaluate pattern-based scores for the text-based context paper set because our 72,000 paper set is small compared to 14 million PubMed papers; thus, a large number of contexts ended up having few direct annotations with training (evidence) papers in our database.
- *Pattern-based Context Paper Set* was created by using a simplified version of the pattern-extraction technique [2]. In this version, only middle tuples of patterns were considered during pattern matching, extended patterns

were not used, and descendant context's papers were included with the ancestor context. If the context contained zero papers, then the closest ancestor's paper set was assigned to the context. Since the ancestor of a context is more general (i.e., less *informative*) than the context itself, assigning papers from an ancestor context to its descendant context introduces a *decay of informativeness* for the context term. A context term's informativeness is approximated through its information content (*I(C)*), which is defined as [13]:

$$I(C) = log(1 / p(C)),$$

where p(C), the relative size of C in GO, is computed as:

$$p(C) = (\# \text{ of descendants of } C) / (\# \text{ of terms in } GO).$$

In order to quantify the *rate of decay*, we compare *I(C)* of the descendant term *(C_{desc})* to that of its ancestor *(C_{ancs})*, and adjust the papers' scores. *RateOfDecay* is defined as:

$$RateOfDecay(C_{ancs}, C_{desc}) = I(C_{ancs})/ I(C_{desc})$$

Citation-based and (simplified) pattern-based scores were assigned to papers in all contexts. Since there are no representative papers defined for the pattern-based context paper set, text-based scores were assigned to only 5,632 contexts that contain at least one representative paper used in the text-based context paper set.

5. Experimental Results

This section evaluates accuracy and separability of the three different score functions.

5.1. Accuracy

We first compare average and median precision scores of selected search terms for both context paper sets. Then, we compare the average top-k overlapping ratio between each pair of score functions over contexts at different context levels to see the changes in accuracy as we drill down in the context hierarchy.

Approximately 120 search terms were used to evaluate the precision scores. These terms were selected from non-GO concepts of external life sciences (genomics) classification systems (e.g., TIGR [23] roles), which have been manually mapped to GO terms [3]. The steps to perform the context-based search in these experiments [2] are: 1) select contexts automatically based on the search term, 2) search within selected contexts, and 3) merge search results from different contexts into a single result set. Only text- and citation-based scores are evaluated for the text-based context paper set because pattern-based scores were not created (as described in section 4). Only pattern- and citation-based scores are evaluated for the pattern-based context paper set because text-based scores were not assigned to all contexts.

Figures 5.1 and 5.2 illustrate the experimental results. We compare precision scores of search results with relevancy scores above various thresholds *t*. When *t* is high, we expect higher precision, which indicates that the search outputs are ranked effectively, i.e., papers receiving high scores are in the true answer set of the search.

Observations:

- Precision scores of the text-based function are higher (> 20%) than those of the citation-based function at moderate thresholds *t* for the text-based context paper set. For the pattern-based context paper set, precision scores of the pattern-based score function are about 10% higher than the citation-based function when *t* is above 0.2.
- When *t* is high, some search terms return no search results. Thus, precisions of these queries are 0, which reduces the average precision scores of all queries. This explains why the average precisions at high *t* (> 0.3) decrease. For these cases, the median precision curves provide a better illustration.
- At high *t*, we observe very high median precision scores. This confirms that the context-based search accuracy proportionally increases with paper ranking.

With respect to precision scores, the citation-based score functions are less accurate than other score functions. Since papers of some contexts cite or are cited by a number of papers outside their contexts, the citation graphs of those contexts are sparse. This causes the citation-based scores to not be highly accurate. Another possible cause is that citations may carry weak indications of topical similarity between citing and cited papers [24], i.e., some citations within a context do not indicate that the citing/cited papers are topically related to the context.

Next, we use the top-k overlapping ratio, defined in section 2, to observe the accuracy changes as we drill down in the context hierarchy. Approximately 5,600 contexts with text-based scores are involved in the experiments. The text-based context paper set is not used because the pattern-based scores were not created. We rely on the top k% as opposed to the top k because the number of lower-level context papers can be significantly smaller than upper-level context papers. Thus, using an absolute k value unfairly biases lower-level contexts. Figure 5.3 illustrates the results.

Observations:

- Overlap ratios between these pairs of score functions decrease as the context level increases: text-based and citation-based, and citation-based and pattern-based. As we drill down in the hierarchy, context size decreases and citation graph sparseness increases. As a result, citation-based scores are less accurate and disagree with other scores in higher-level contexts.

- The text-based and the pattern-based scores agree less with each other when the contexts are closer to the root level. As the selected context nears the root level, the context becomes more general and includes a large number of subcontexts. For text-based prestige score computation, representative papers of more general contexts may not characterize the context very well. For pattern-based prestige score computation, the building blocks of patterns (i.e., significant terms) become less selective for more general contexts and may result in incorrect scores.

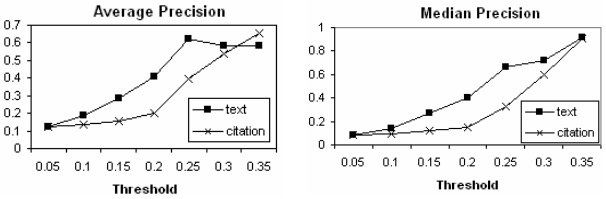


Fig. 5.1. Precision scores of text-based context paper set

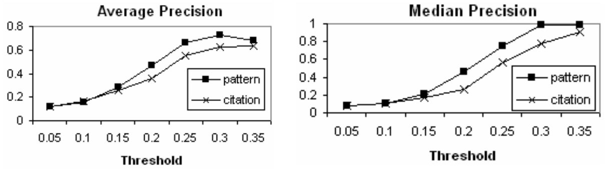
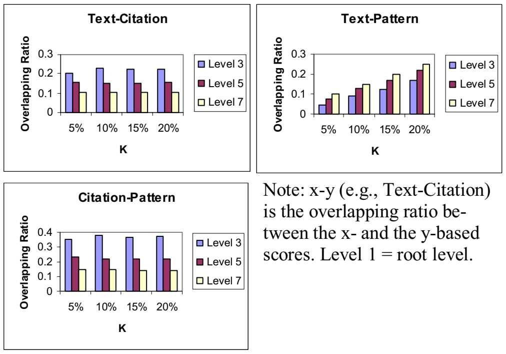


Fig. 5.2. Precision scores of pattern-based context paper set



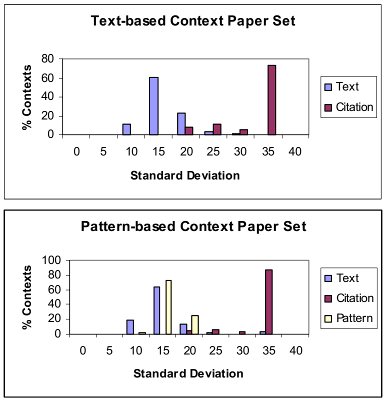
Note: x-y (e.g., Text-Citation) is the overlapping ratio between the x- and the y-based scores. Level 1 = root level.

Fig. 5.3. Average top-k% overlapping ratio per context level

5.2. Separability

The best separability for a function *f* occurs when *f* uniformly maps the points of its input domain to those of its output domain. Assuming scores are divided into *k* ranges for each context, the percentage of papers with scores in each range should be (100/k)%. E.g., Assume papers in every context c_i receive scores between [0, 1], and scores are divided into 10 ranges of [s, s+0.1] where $0 \leq s \leq 0.9$. For a score function to have perfect separability, 10% of papers in c_i should receive the prestige scores in each range. If the score function possesses good separability, the number of papers with scores in each score range should be almost equal, and the standard deviation should approach 0. The standard deviation (SD) used in the experiments for context c_i is defined as: $SD=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}$, where *n* is the number of score ranges, $1 \leq i \leq n$, X_i is the percentage of papers in context c_i with scores in range $\left[\frac{1}{n}*(i-1),\frac{1}{n}*i\right]$, and $\overline{X} = 100/n$.

We first evaluate the separability of score functions in all selected contexts. Then, we show the changes in separability as we drill down in the context hierarchy. Figure 5.4 illustrates the overall score distributions. If a score function possesses good separability, we expect a large number of contexts to have low standard deviations.



Note: pattern, citation, and text refer to pattern-, citation-, and text-based scores, respectively

Fig. 5.4. Histogram of percentage of contexts by standard deviation

Observations:

- The overall separability of the citation-based score function is worse than the text-based and the pattern-based functions. Since the citation graphs of many contexts are sparse, the PageRank algorithm assigns a small number of unique paper scores in those contexts.
- The text-based score function yields the best separability. As shown in figure 5.4, the standard deviations of most contexts are quite low (< 15) for the text-based scores.
- Both text-based and pattern-based score distributions

are closer to normal (in addition to uniform) versus the citation-based score distribution. For the text-based and the pattern-based functions, the majority of contexts show moderate levels of deviation. On the other hand, most contexts for the citation-based function show very high deviation.

Figures 5.5 - 5.7 illustrate the score distribution at each context level for text-based, pattern-based, and citation-based score functions, respectively.

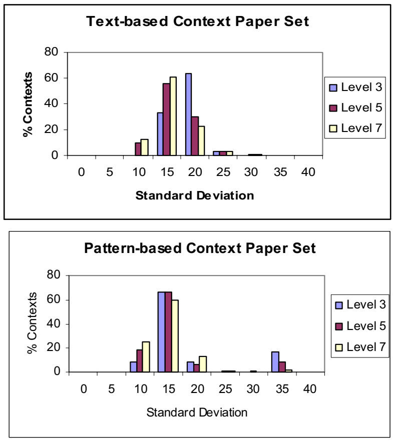


Fig. 5.5. Score distribution per context levels for "text-based scores"

Observation: As we drill down in the context hierarchy, the separability of the text-based scores increases. In figure 5.5, the percentage of contexts at level 7 with low standard deviations (< 10) is higher than levels 3 and 5.

When contexts are closer to the root, the contexts' paper sets are larger and more diverse. Therefore, it is harder to find representative papers that accurately characterize the upper-level contexts. Thus, most of the papers in the upper-level contexts receive small and not well-distributed text-based scores compared to the lower-level contexts.

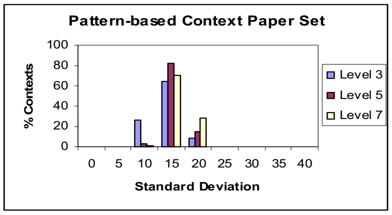


Fig. 5.6. Score distribution per context levels for "pattern-based scores"

Observation: The pattern-based score separability of a context *C* is inversely proportional to *C*'s context level.

To illustrate the above observation, we give the following example. GO term "RNA polymerase II transcription factor activity", which we call X, has four children: "general X", "nonspecific X", "X, enhancer bin", and "specific X". X has several siblings, e.g., "transcription cofactor activity", "transcription elongation regulator activity", etc. It is easier to distinguish between X's siblings than X's children since the number of different words between siblings is higher. Also, the context terms closer to the root level become more general. As a result, the number of constructed patterns in the parent contexts tends to be higher than their child contexts. Since each constructed pattern has its own pattern score, and more patterns potentially result in more matches, there is a greater chance that the paper scores of upper-level contexts are more diversified.

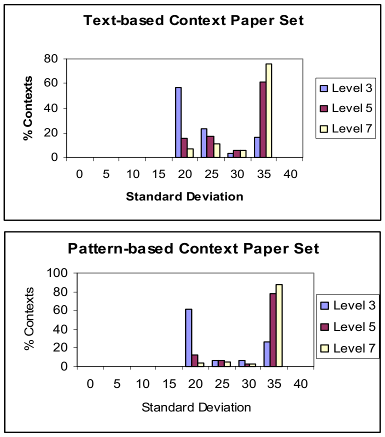


Fig. 5.7. Score distribution per context levels for "citation-based scores"

Observation: The citation-based score separability of a context *C* is inversely proportional to *C*'s context level.

As we drill down in the context hierarchy, the citation graphs of the contexts are sparser. This causes the PageRank algorithm to assign a small number of unique paper scores for these contexts. As a result, citation-based scores have low separability.

6. Related Work

While many literature search systems are available online, only GoPubMed [22] uses context hierarchies. GoPubMed queries are submitted to PubMed, and the corresponding PubMed paper "abstracts" are retrieved and categorized by GO terms. However, categorization fully relies on the existence of GO term words *in the abstracts*, and only 78% of the 14 million PubMed abstracts contain words occurring in a GO term (as seen by using our *PubMed Abstracts FullText Search Tool* [21]). GoPubMed does not rank results or provide importance scores for papers.

Several contextual web search approaches aim to improve keyword-based search accuracy. In one approach, a context is captured around the user-highlighted text, and augmented queries are created from the selected context words [16, 18]. This approach relies on user-defined contexts and uses no hierarchical structure. Another approach clusters search results into automatically-derived hierarchical contexts [25]. While the constructed contexts are closely related to the search results, they are not as meaningful as the human-created ontology-based contexts like GO. Also, users cannot select contexts of interest before viewing search results or modify search results beyond the constructed contexts. Topic Sensitive PageRank [17] creates 16 topic-sensitive PageRank vectors with each vector biased by URLs in the top level of the Open Directory Project [19]. The citation-based prestige score function presented in this paper is similar to the Topic Sensitive PageRank, but we consider more specific (non-top-level) contexts in the context hierarchy.

7. Conclusion and Future Work

We presented three different prestige score functions for ranking papers in a context-based environment, namely, citation-based, text-based, and pattern-based score functions. We evaluated the quality of a score function by measuring its accuracy and separability. For the context-based environment, we showed that the text-based and the pattern-based scores yield better accuracy and separability than the citation-based scores.

A possible future work is to add a variation on score function computations. Instead of omitting relationships from different contexts during prestige score computations, we can assign weights to these relationships. In other words, author and citation relationships from other contexts can boost paper scores in a context c_1. E.g., if a paper p_a of context c_2 cites or is cited by a paper in c_1, the citation relationship weight of p_a in c_1 may be assigned as follows. If c_2 is not hierarchically related to c_1 (i.e., c_2 is not a close relative of c_1), assign the smallest weight. If c_2 is hierarchically related to c_1, assign a higher weight. If p_a is in c_1, assign the highest weight.

References

[1] PubMed, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[2] Ratprasartporn, N., Po, J., Cakmak, A., Bani-Ahmad, S., Ozsoyoglu, G., “On Context-Based Publication Search Paradigm: Gene-Ontology-Specific Contexts for Searching PubMed Effectively”. Technical Report, CWRU 2006

[3] Gene Ontology, http://geneontology.org/

[4] Cakmak, A., Ozsoyoglu, G., “Annotating Genes Using Textual Patterns”. PSB 2007

[5] Agrawal, R. and Ramakrishnan, S. “Fast Algorithms for Mining Association Rules”. VLDB 1994.

[6] Salton, G., “Automatic Text Processing”, Addison-Wesley, 1989.

[7] Al-Hamdani, A., “Querying Web Resources with MetaData in a Database”. PHD Dissertation CWRU, 2004

[8] Brin, S., Page, L., “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Computer Networks and ISDN Systems, 1998.

[9] Kleinberg, J. M., “Authoritative Sources in a Hyperlinked Environment”, ACM-SIAM Symp. on Discr Alg., 1998

[10] Cakmak, A.,"HITS- and PageRank-based Importance Score Computations for ACM Anthology Papers",Tech. Report, CWRU, 2003

[11] Bani-Ahmad, S., Cakmak, A., Al-Hamdani, A., Ozsoyoglu, G., “Evaluating Score and Publication Similarity Functions in Digital Libraries”, Technical Report, CWRU 2005

[12] Po, J., “Context-Based Search in Literature Digital Libraries”, MS Thesis, CWRU 2006.

[13] Resnik, P., “Using Information Content to Evaluate Semantic Similarity in a Taxonomy”. IJCAI, 1995.

[14] Small, H., “Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents”, Journal of the American Society for Information Science, Vol. 24, No. 4 28-31, 1973.

[15] Kessler, M. M., “Bibliographic Coupling between Scientific Papers, American Documentation”, 14:10-25, 1963.

[16] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E., “Placing Search in Context: The Concept Revisited”, WWW 2001.

[17] Haveliwala, T.H., “Topic Sensitive PageRank”, WWW 2002

[18] Kraft, R., Chang, C.C., Maghoul, F., Kumar, R., “Searching with Context”, WWW 2006.

[19] The Open Directory Project, http://www.dmoz.org/

[20] Chakrabarti, S., “Mining the Web, Discovering Knowledge from Hypertext Data”, Morgan-Kaufmann, 2003

[21] PubMed Abstracts FullText Search, http://nashua.case.edu/PubmedExplorer/index.aspx.

[22] Delfs, R., Doms, A., Kozlenkov, A., Schroeder, M., “GoPubMed: ontology-based literature search applied to Gene Ontology and PubMed”, German Conference on Bioinformatics 2004: 169-178.

[23] The Institute for Genomic Research (TIGR), http://www.tigr.org/

[24] Aya, S., Lagoze, C., Joachims, T., “Citation Classification and its Applications”, ICKM 2005.

[25] Ferragina, P., Gulli, A., “A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering”, WWW 2005.

[26] ACM Digital Library, http://www.acm.org/dl

[27] Google Scholar, http://scholar.google.com