Combining Multiple Sources of Evidence to Enhance Web Search Performance

Kiduk Yang*

〈Contents〉

Ⅰ. Introduction
Ⅱ. Previous Research
 1. Link Analysis
 2. Fusion IR
Ⅲ. Experiment
 1. Data
 2. Methodology
Ⅳ. Results
 1. Single System Results
 2. Fusion Results
 3. Overlap Analysis
Ⅴ. Concluding Remarks

ABSTRACT

The Web is rich with various sources of information that go beyond the contents of documents, such as hyperlinks and manually classified directories of Web documents such as Yahoo. This research extends past fusion IR studies, which have repeatedly shown that combining multiple sources of evidence (i.e. fusion) can improve retrieval performance, by investigating the effects of combining three distinct retrieval approaches for Web IR: the text-based approach that leverages document texts, the link-based approach that leverages hyperlinks, and the classification-based approach that leverages Yahoo categories. Retrieval results of text-, link-, and classification-based methods were combined using variations of the linear combination formula to produce fusion results, which were compared to individual retrieval results using traditional retrieval evaluation metrics. Fusion results were also examined to ascertain the significance of overlap (i.e. the number of systems that retrieve a document) in fusion. The analysis of results suggests that the solution spaces of text-, link-, and classification-based retrieval methods are diverse enough for fusion to be beneficial while revealing important characteristics of the fusion environment, such as effects of system parameters and relationship between overlap, document ranking and relevance.

Keywords: Fusion, Web search, Information retrieval

초 록

웹은 하이퍼링크 및 야후와 같이 수동으로 분류된 웹 디렉토리 처럼 문서의 콘텐츠를 넘어선 다양한 정보의 소스가 풍부하다. 이 연구는 웹문서 내용을 활용한 텍스트기반의 검색 방식, 하이퍼 링크를 활용한 링크 기반의 검색 방식, 그리고 야후의 카테고리를 활용한 분류 기반의 검색 방식을 융합하므로서 여러 정보소스를 결합하면 검색 성능을 향상시킬 수 있다는 기존 융합검색연구들을 확장시켰다. 텍스트, 링크, 분류 기반 검색 결과를 여러가지 선형조합식으로 생성한 융합결과를 기존의 검색 평가 지표를 사용하여 각각의 검색 결과와 비교 한 후, 검색결과 오버랩의 중요성 또한 조사 하였다. 본 연구는 텍스트, 링크, 분류 기반 검색의 솔루션 스페이스들의 다양성이 융합검색의 적합성을 제시한다는 결론과 더불어 시스템 파라미터의 영향, 그리고 오버랩, 문서순위, 관련성들의 상호 관계 같은 융합 환경의 중요한 특성들을 분석하였다.

키워드: 융합, 웹검색, 정보검색

* Associate Professor, Department of Library and Information Science, Kyungpook National University (kiyang@knu.ac.kr)

· 논문접수: 2014년 8월 25일 ·최초심사: 2014년 8월 25일 ·게재확정: 2014년 9월 12일

Ⅰ. Introduction

The Web document collection is rich in sources of evidence(e.g., text, hyperlinks, Web directories) and thus offers an opportunity to employ a diverse set of retrieval approaches that can leverage a wide range of information. Furthermore, findings from fusion information retrieval (IR) research suggest that multiple sources of evidence(MSE) and multiple methods that leverage them could be combined to enrich the retrieval process on the Web. The nature of the Web search environment is such that retrieval approaches based on single sources of evidence suffer from weaknesses that can diminish the retrieval performance in certain situations. For example, text-based IR approaches have difficulty dealing with the diversity in vocabulary and quality of web documents, while link-based approaches can suffer from incomplete or noisy link topology. The inadequacies of singular Web IR approaches coupled with the fusion hypothesis of "fusion is good for IR" make a strong argument for combining MSE as a potentially advantageous retrieval strategy for Web search.

The first step in fusion IR is to identify and examine the sources of evidence to combine. MSE on the Web that can be leveraged for IR are: contents of Web documents, characteristics of Web documents, hyperlinks, Web directories such as Yahoo, and user statistics. The main hypothesis of the study, namely that combining text-, link-, and classification-based[1)] methods can enhance Web search performance, is based on three observations. First, each of text-, link-, and classification-based approaches suffers from individual weaknesses that hinder optimum retrieval performance. Second, all three approaches have complementary strengths that can boost retrieval performance when combined. Third, there is ample evidence in literature that suggests fusion to be beneficial for IR.

Text-based approaches have difficulties dealing with the vocabulary problem(e.g., different expressions of the same concept in Web documents and queries), the diversity of document quality and content, fragmented documents, and documents with little textual contents, such as "hubs", index pages, and bookmarks. Link-based approaches do not fare well when faced with a variety of link types(e.g., citation links, navigational links, commercial links, spam links), and

1) Henceforth, classification-based method will refer to a method that leverages Web directory information.

"emerging" communities with incomplete link topologies. Web directories, in addition to classification and vocabulary problems(e.g., different categorizations of the same Web document and different labeling of the same category), contain only a fraction of the documents on the Web.

The most obvious of the complementary strengths is found in the combination of text- and link-based approaches. Ranking text-based retrieval results by a measure "link importance" can help differentiate documents with similar textual contents but varying degrees of importance, popularity, or quality. Pruning documents based on their textual contents before applying link analysis techniques can help alleviate the problems introduced by spurious links. On an abstract level, web directories, which embody explicit human judgments about the quality and topicality of documents, can augment ranking algorithms based on counting of words or links. Specifically, Web directories can be used to train document classifiers, to find related documents, and to help refine queries by finding subcategories, searching within a category, and suggesting related categories, all of which employ text- and/or link-based methods to exploit the information contained in Web directories.

Combining text-, link-, and classification-based methods can be viewed as attempting to maximize the combined strengths of leveraging author's knowledge, peer's knowledge, and cataloger's knowledge about Web documents while minimizing their individual weaknesses. Although the idea of fusion is intuitively appealing, there is a shortage of techniques that utilize the considerable body of explicit human judgment(e.g. Web directories[2]) in combination with hyperlinks and textual contents of Web documents. IR research dealing with knowledge organization focus mostly on automatic clustering and classification of documents, and there is little investigation on how existing hierarchical knowledge bases like Yahoo can be brought into the fusion of text retrieval and link analysis techniques in Web IR. Consequently, this work explores the question of combining link analysis, content analysis, and classification-based techniques to improve retrieval performance.

2) Web directories will refer to manually constructed topical taxonomies of Web documents(e.g. Yahoo) in this paper.

Ⅱ. Previous Research

1. Link Analysis

Two best-known link analysis methods, which are based on the notion that hyperlinks contain implicit recommendations about the documents to which they point, are Page Rank and HITS (Yang 2005).

Page Rank is a method for assigning a universal rank to Web pages based on a recursive weight-propagation algorithm(Page et al. 1998). Page et al. start with the notion of counting backlinks(i.e., indegree) to assess the importance of a Web page, but point out that simple indegree does not always correspond to importance; thus they arrive at propagation of importance through links, where a page is important if the sum of the importance of its backlinks is high.

Kleinberg's(1999) HITS(Hyperlink Induced Topic Search) algorithm considers both inlinks and outlinks to identify mutually reinforcing communities of "authority" and "hub" pages. Though HITS embraces the link analysis maxim that says a hyperlink is an annotation of human judgment conferring authority to pointed pages, it differs from other link-based approaches in several regards. Instead of simply counting the number of links, HITS calculates the value of page *p* based on the aggregate values of pages that point to *p* or are pointed to by *p*, much in the same fashion as Page Rank. HITS, however, differs from Page Rank in three major points. First, it takes into account the contributions from both inlinks and outlinks to compute two separate measures of a page's value, namely authority and hub scores, instead of the single measure of importance like Page Rank. Second, HITS measures pages' values dynamically for each query, rather than assigning their global scores once and for all regardless of any query. Third, HITS scores are computed from a relatively small subset of the Web instead of the totality of the Web.

2. Fusion IR

The bulk of IR fusion research, which investigates the various ways of combining different retrieval strategies, have found fusion to have a positive effect on retrieval performance regardless of what strategies were combined. The potential of fusion to leverage the strengths of its

components while minimizing their weaknesses is not only promising in its own right, but offers a novel perspective of IR that relaxes the research goal of discovering the one best retrieval strategy.

Earlier studies on combining different document representations discovered that combined representations achieved better retrieval outcome than single representations(e.g., title and keyword vs. title or keyword) despite the fact that the difference in retrieval performance among single representations were small(Keen 1973; Spark Jones 1974). To explain this phenomenon, subsequent studies examined the overlap between different document representations(i.e., common terms) and found overlap to be small(Williams 1977; Smith 1979). A more systematic study of different document representations' relative effectiveness was later conducted by Katzer et al.(1982), who executed 84 queries on seven representations of 12,000 INSPEC documents and found but higher overlap among relevant documents than nonrelevant documents. The relationship between overlap and relevance was further studied from the perspective of the searcher by Saracevic and Kantor(1988), who found that the odds of a document being relevant increased monotonically with the number of retrieved sets in which it appeared. Fox and Shaw(1994; 1995), in combining different retrieval methods as well as query representations, discovered that combining dissimilar sources of evidence was better than combining similar ones.

Lee(1996) experimented with the fusion of multiple relevance feedback methods and found that different relevance feedback methods retrieve different sets of documents even though they achieve a similar level of retrieval effectiveness. By examining the overlap of relevant and non relevant documents retrieved by different retrieval systems separately, Lee(1997) observed what he calls the “unequal overlap property”, which is the phenomenon that different systems tend to retrieve similar relevant documents but dissimilar nonrelevant documents. The unequal overlap property challenges the fusion assumption by Belkin et al.(1993), which suggests that retrieval performance improvement by fusion may be partially due to combining of different relevant documents retrieved by different sources of evidence. In fact, Lee hypothesized that fusion is warranted when systems retrieve similar sets of relevant documents but different sets of nonrelevant documents. Lee’s hypothesis about the condition for effective fusion was more formally validated by Vogt and Cottrell(1998). Based on the performance estimation of a linearly combined system based on measurable characteristics of the component systems, they concluded that to achieve effective fusion by linear combination, one should maximize both the individual

system's performance and the overlap of relevant documents between systems, while minimizing the overlap of nonrelevant documents. They also suggested that component systems should distribute scores similarly but not rank relevant documents similarly.

We should note at this point that the art of fusion lies in discovering how different experts or sources of evidence can be combined to exploit their strengths while at the same time remaining unaffected by their weaknesses. As the study by Bartell et al.(1994) demonstrates, the fusion solution space is defined not only by the fusion algorithm but also by how that fusion algorithm is applied. In other words, the optimal fusion approach should consider the context of fusion in determining its overall fusion strategy.

Ⅲ. Experiment

This study combines methods that leverage text, hyperlink, and Web directory information to improve retrieval performance. Retrieval results of methods that leverage Web page text, hyperlink, and Yahoo Web directory information, both combined and individual, were examined to determine whether such fusion approaches are beneficial for Web search. The following sections describe the experiment in more detail.

1. Data

The data used for the study is the WT10g test collection from Text REtreival Conference (http://trec.nist.gov/), which consists of 1.7 million Web documents, 100 queries(topics 451-550), and relevance judgments. The WT10g collection also includes the connectivity data, which provides lists of inlinks and outlinks of all documents in the collection. The connectivity data, however, is restricted to the WT10g universe, meaning that inlinks and outlinks that are not part of the collection themselves are omitted, which is likely to make the link topology of the collection incomplete. The document set of WT10g is comprised of 1,692,096 html pages, arrived at by sampling the original Internet Archive data in such a way to include a balanced representation of the real Web characteristics such as hyperlink structure and content types. Duplicates, non-English, and binary documents were excluded from the collection. Queries, or

topics as they are called in TREC, consist of a title field, containing actual queries as they were submitted to Web search engines, a description field, which is typically a one sentence description of the topic area, and a narrative field, containing descriptions of what makes documents relevant. The description and narrative fields were created by TREC assessors to fit the intent of the real Web search queries represented in the title.

The characteristics of a Web directory, such as breadth of coverage, consistency of classification, and granularity of categories, are important factors to consider in determining the data source for classification-based retrieval methods. An ideal Web directory would be one that has classified all the documents of the test collection into fine-grained categories in a consistent manner. Lacking the ideal Web directory, Yahoo(http://yahoo.com) was used to leverage Web directory information for its size and popularity. Instead of using the actual Web documents associated with Yahoo categories, the classification-based method uses document titles and Yahoo's descriptions of the categorized pages to represent each categorized document. This not only speeds up processing but also reduces noise in representation, which is a preferred practice in text categorization. In addition, the annotated description of a Yahoo site, along with the category to which it belongs, represent the cataloger's knowledge about the classified document, which complements the author's knowledge embodied in the document text and peers' knowledge embodied in hyperlinks that point to the document.

2. Methodology

2.1 Text-based Retrieval Method

The text-based retrieval component is based on a Vector Space Model(VSM) using the SMART length-normalized term weights (Singhal et al. 1996). Documents are processed by first removing markup tags and punctuation and then excluding stopwords, low frequency terms, non-alphabetical words(exception: embedded hyphen), words consisting of more than 25 or less than 3 characters, and words that contain 3 or more repeated characters. After punctuation and stopword removal, each word was conflated by applying the simple plural remover(Frakesand Baeza-Yates 1992). The simple plural remover was chosen to speed up indexing time and to minimize the overstemming effect of more aggressive stemmers. TREC topics were stopped and stemmed in the same manner as the document texts.

In addition to body text terms(i.e. terms between <BODY> and </BODY> tags), header text terms were extracted from document titles, meta-keyword and description texts, and heading texts (i.e. texts between <Hn> and </Hn> tags). A combination of body and header text terms was also created, where the header text terms were emphasized by multiplying the term frequencies by 10.In each of the three term sources(i.e. body, header, body and header), noun phrases were identified to construct noun phrase indexes. A noun phrase is defined as up to three adjacent nouns or capitalized words within a phrase window.

The VSM method used SMART *Lnu* weights for document terms(Buckley et al. 1996; 1997) and SMART *ltc* weights (Buckley et al. 1995) for query terms. *Lnu* weights attempt to match the probability of retrieval given a document length with the probability of relevance given that length (Singhal et al. 1996).The formula for *Lnu* document term weight is:

$$d_{ik} = \frac{(1+\log(f_{ik}))/(1+\log(avg\_f_i))}{(1.0-slope)*p_i + slope*T}, \quad (1)$$

where f_{ik} is the number of times term k appears in document i(i.e. in-document frequency), avg_f_i is the average in-document frequency for document i, T is the number of unique terms in the collection, p_i is the average number of unique terms in a document i, and *slope* is the normalization parameter that can be adjusted for a given document collection.The formula for *ltc* query term weight is:

$$q_k = \frac{(\log(f_k)+1)*idf_k}{\sqrt{\sum_{j=1}^{t}\left[(\log(f_j)+1)*idf_j\right]^2}}, \quad (2)$$

where f_k is the number of times term k appears in the query, and idf_k is the inverse document frequency of term k. The denominator is a document length normalization factor, which compensates for the length variation in queries.

In addition to initial retrieval results, top ten positive and top two negative weighted terms from the feedback vector, created by the adaptive linear model using the top three ranked documents of the initial retrieval result as “pseudo-relevant”, were used as a query to generate the

pseudo-feedback retrieval results. The basic approach of the adaptive linear model, which is based on the concept of preference relations from decision theory(Fishburn 1970), is to find a *solution vector* that will rank a more-preferred document before a less-preferred one(Wong et al. 1988). The solution vector is arrived at via an *error-correction procedure*, which begins with a starting vector $\mathbf{q}_{(0)}$ and repeats the cycle of "error-correction" until a vector is found that ranks documents according to the preference order estimation based on relevance feedback(Wong et al. 1991). The error-correction cycle i is defined by

$$\mathbf{q}_{(i+1)} = \mathbf{q}_{(i)} + \alpha\mathbf{b} \ , \qquad (3)$$

wherea is a constant, and **b** is the *difference vector* resulting from subtracting a less-preferred document vector from a more preferred one(Sumner et al. 1998).

Table 1 enumerates the text-based method parameters for VSM systems, which are query length, term source, use of phrase terms, and use of pseudo-feedback. Query length ranges from short(topic title) and medium(topic title and description) to long(topic title, description, and narrative). Term sources are body text, header text, and body plus header text. The use of noun phrase indicates whether the term index for each term source contains both single and phrase terms or single terms only. The combination of parameters(3 query lengths, 3 term sources, 2 for phrase use, 2 for feedback use) resulted in 36 VSM systems.

<Tab. 1> VSM system* parameters

Query Length	*Term Source*	*Noun Phrase*	*Pseudo-feedback*
short (*s*)	body text (*b*)	no (*0*)	no (*1*)
medium (*m*)	header text (*h*)	yes (*1*)	yes (*2*)
long (*l*)	body + header (*bh*)		

*VSM system name = vsm$qform$index$phrase.$feedback(e.g. vsmsb0.1)
where $qform=*Query Length*, $index=*Term Source*, $phrase=*Noun Phrase*, $feedback=*Pseudo-feedback*

2.2 Link-based Retrieval Method

For the study, the authority score of documents computed by the HITS algorithm (Kleinberg 1999) is used to generate a ranked list of documents with respect to a given query. HITS was chosen over PageRank scores, whose effective computation requires a link propagation over much larger set of linked documents than the WT10g corpus(Brinand Page 1998), and the Clever algorithm(Chakrabarti

et al. 1998), which makes it difficult to isolate the contributions and behaviors of individual methods by implicitly combining link- and text-based methods to extend HITS.

HITS defines "authority" as a page that is pointed to by many good hubs and defines "hub" as a page that points to many good authorities. Mathematically, these circular definitions can be expressed as follows:

$$a(p) = \sum_{q \to p} h(q) \quad , \quad h(p) = \sum_{p \to q} a(q) \tag{4}$$

The above equations define the authority weight *a(p)* and the hub weight *h(p)* for each page p, where $p \rightarrow q$ denotes "page *p* has a hyperlink to page *q*". HITS starts with a root set *S* of text-based search engine results in response to a query about some topic, expands *S* to a base set *T* with the inlinks and outlinks of *S*, eliminates links between pages with the same domain name in *T* to define the graph *G*, runs an iterative algorithm on *G* until convergence, and returns a set of documents with high *h(p)* weights(i.e. hubs) and another set with high *a(p)* weights(i.e. authorities).

The original HITS algorithm was modified by adopting a couple of improvements from other HITS-based approaches. As implemented in the ARC algorithm (Chakrabarti et al. 1998), the root set was expanded by 2 links instead of 1 link(i.e. expand *S* by all pages that are 2 link distance away from *S*). Also, the edge weights by Bharat and Henzinger(1998), which essentially normalize the contribution of authorship by dividing the contribution of each page by the number of pages created by the same author, was used to modify the HITS formulas as follows:

$$a(p) = \sum_{q \to p} h(q) \times auth\_wt(q,p) \quad , \quad h(p) = \sum_{p \to q} a(q) \times hub\_wt(p,q) \quad . \tag{5}$$

In above equations, *auth_wt(q,p)* is $1/m$ for page *q*, whose host has m documents pointing to *p*, and *hub_wt(p,q)* is $1/n$ for page *q*, which is pointed by *n* documents from the host of *p*. To establish a definition of a host, we opted for a simplistic method based on URL lengths. Short host form was arrived at by truncating the document URL at the first occurrence of a slash mark (i.e. '/'), and long host form from the last occurrence.

〈Tab. 2〉 HITS system* parameters

Host Definition	*Seed Set*
short (*s*)	short query, body text, phrase, no feedback (*sb1.1*)
long (*l*)	medium query, body text, phrase, no feedback (*mb1.1*)
	long query, body text, phrase, no feedback (*lb1.1*)

*HITSsystem name = hit$hform$seed(e.g. hitssb1.1)
where$hform = *Host Definition*, $seed = *Seed Set*

Among the 36 text-based system results, we chose the best performing system with all variations of query lengths as the seed sets. The combination of host definition and seed set parameters, as seen in Table 2, resulted in 6 HITS systems.

2.3 Classification-based Retrieval Method

The first step of the classification-based method is to find the categories that best match the query, which can be thought of as a query-category matching problem. Since Web directories classify only a fraction of the Web and do not rank documents that are classified, a second step that classifies and ranks documents with respect to the best matching category is needed, which can be thought of as a category-document matching problem.

To leverage the classification information, we employed the *Term Match*(TM) method that selects the best Yahoo categories for a query based on matching of query-category terms. The first step of the TM method matches query terms to terms in the Yahoo categories, which we reextracted from category labels, site titles, descriptions and URLs, to generate a ranked list of categories. In the second step, an expanded query vector is constructed from the best matching categories to produce a ranked list of the WT10g documents. The expanded query vector consists of terms in the original query, the label of the best matching category, and the titles and descriptions of the top three sites[3] in the best matching categories. The term weight of the expanded query vector, which is computed by multiplying the term-category association weight by term frequency and dividing by the vector length, can be expressed as:

$$q_{kc} = \frac{f'_k * cd_{kc}}{\sqrt{\sum_{j=1}^{t} (f'_j * cd_{jc})^2}}, \qquad (6)$$

3) Sites in a matching category are ranked by the number of unique query terms in their titles and descriptions.

where cd_{kc} is the association weight of term k to category c, f'_k is the total number of times term k appears in the query, the label of category c, and the titles and descriptions of the top three sites of category c, and the denominator is the length-normalization factor. The term association weight, which estimates the probability of association between terms and categories (Plauntand Norgard 1998),is computed by the formula below[4]:

$$\lambda' = 2\left[\log L(p_1,k_1,n_1) + \log L(p_2,k_2,n_2) - \log L(p,k_1,n_1) - \log L(p,k_2,n_2)\right], \quad (7)$$

where:

$$\log L(p,k,n) = k\log p + (n-k)\log(1-p),$$

$$p_1 = \frac{k_1}{n_1}, \quad p_2 = \frac{k_2}{n_2}, \quad p = \frac{k_1+k_2}{n_1+n_2}, k_1 = AB, \quad n_1 = AB+\neg AB, \quad k_2 = A\neg B, \text{ and } n_2 = A\neg B+\neg A\neg B.$$

〈Tab. 3〉 TM system* parameters

# of top categories	*WT10g index*	*Pseudo-feedback*
1	body text, no phrase (*b0*)	no (*1*)
2	body text, phrase (*b1*)	yes (*2*)
3	body+header, no phrase (*bh0*)	
	body+header, phrase (*bh1*)	

*TMsystem name =tmcnwindex.$feedback (e.g. tm1b0.1)
where$cn = # of top categories, $windex = WT10g index, $feedback = pseudo-feedback

The TM method finds the best category for a query based on the number of matching query terms, and expands the original query with selected category terms that are weighted by term association weights to rank documents. The way the TM method leverages the classification information is twofold. First, it uses manually assigned category terms(i.e., category labels, site titles and descriptions) to find the best matching categories and to expand the query. Second, it uses term association weights, which are based on term-category co-occurrence, to compute the term weights of the expanded query vector. In other words, the importance of category labels as well as multi-term concepts are underscored in ranking categories by the number of unique query terms in category labels, while the importance of term and category co-occurrence in the classification hierarchy is emphasized in the term weights of the expanded query vector.

4) A contingency table containing the counts of each of the possible combinations of term *A* and category *B* is constructed, where "Ø" denotes the absence of some event.

The parameters tested for the TM systems are number of top (i.e., best matching) categories used, WT10g term index, and use of pseudo-feedback. The combination of the parameters (3 top categories, 4 WT10g term index, 2 for feedback use) resulted in 24TM systems (Table 3).

2.4 Fusion Method

How to combine or integrate fusion components is the key question in fusion IR. One of the most common fusion methods is the *Weighted Sum*(WS) formula Bartell et al. 1994; Modhaand Spangler 2000), which computes the fusion score by the weighted sum of individual retrieval scores. When fusion component systems are highly distinct from one another, normalization of retrieved documents cores across systems may not be able to compensate for differences in document ranking functions. This is the case with fusion of text-, link-, and classification-based retrieval systems, whose document scores are computed in different manners to measure different values of a document with respect to a query. More specifically, the document score of VSM systems measures the textual similarity between a query and a document, the document score of HITS systems represents the hyperlink-conferred authority of a document for the topic of a query, and the document score of TM systems measures the likelihood of a document belonging to the same category as the query. In such scenarios, it might be useful to merge the ranks of a document instead of the scores of documents.

To compensate for the differences among fusion component systems, the *Weighted Rank Sum* (WRS) formula, which uses rank-based scores(e.g., 1/rank) in place of document scores of the WS formula, was tested:

$$FS = \sum(w_i * RS_i), \quad (8)$$

where: FS = fusion score of a document,

w_i = weight of system i,

RS_i = rank-based score of a document by system I.

Although the WRS formula aims to weight the contributions of individual fusion components to the retrieval outcome by their relative strength, it does not explicitly differentiate between overlapped and non-overlapped instances. In other words, the absolute contribution of a document retrieved by a system remains the same whether the document in question is retrieved by other

systems or not. What WRS formula neglects is the possibility that the contribution of a document by a system to the retrieval outcome could be different across overlap partitions(e.g., documents retrieved by system 1 only, by system 1 and 2 only, etc.).

To leverage the retrieval overlap and rank information, we devised two additional fusion formulas. *Overlap Weighted Rank Sum*(OWRS), attempts to leverage overlap while compensating for the differences among fusion component systems by weighting rank-based scores by overlap partitions(Equation 9). *Rank-Overlap Weighted Rank Sum*(ROWRS) is a variation of the OWRS formula that considers not only the overlap partition but also the rank at which a document is retrieved in its computation of weights(Equation 10).

$$FS = \sum(w_{ik} * RS_i), \quad (9)$$

where: FS = fusion score of a document,
w_{ik} = weight of system i in overlap partition k,
RS_i = rank-based score of a document by system i.

$$FS = \sum(w_{ikj} * RS_i), \quad (10)$$

where: FS = fusion score of a document,
w_{ikj} = weight of system i in overlap partition k at rank j,
RS_i = rank-based score of a document by system i.

In all the formulas of Weighted Sum variation, topics 451 to 500 were used as training data to determine the weights. Overall average precision, which is a single-value measure that reflects the overall performance over all relevant documents, was used to determine the weight in the WRS formula. In the OWRS formula, overall average precision was multiplied by overlap average precision, which is the average precision computed for each overlap partition. In a 3-system fusion, for example, average precision is computed for each of the 4 overlap partitions for each system. In other words, the result set of a system is partitioned into overlap partitions (e.g., for system A: documents retrieved by system A, by system A and B, by system A and C, by system A, B, and C), and average precision is computed in each partition of each system.

For the ROWRS, which needs a performance estimate at a given rank, overall average precision is not appropriate. Instead, three rank-based measures, namely *Precision(P)*, *Effectiveness(F)*, and

Success/Failure(sf) at each rank, were used to compute the weights in three versions of the ROWRS formula. Since weights based on rank-based measures can be overly sensitive to the exact rank of a document, they were applied in "rank blocks"(e.g., ranks 1 to 10, 11 to 20, etc.). In other words, fusion component scores(RS_i in Equation 10) in a given rank block had the same weights, which was determined by averaging rank-based measures over rank blocks.

Whereas *P* and *F* measures are based on performance up to a given rank *k*(e.g., number of relevant documents in the top *k* results), *sf* measure is based on the retrieval success/failure at each rank(e.g., 1/*k* if the document at rank *k* is relevant, 0 otherwise). The *sf* measure estimates the system performance at a given rank interval without regard to its performances in prior rank intervals in an attempt to boost the results of systems that retrieve relevant documents at lower ranks. For example, a non-relevant document at rank 101 with 100 relevant documents in rank 1 to 100(doc-A) will have much higher *P* and *F* than a relevant document at rank 101 with 0 relevant documents in rank 1 to 100(doc-B), but the *sf* of doc-B will be higher than the *sf* of doc-A. When fusion components include systems that retrieve relevant documents in lower-ranked clusters, such an approach might be beneficial. Since HITS systems can retrieve relevant documents in the non-principal communities of hubs and authorities, and the best matching category of TM systems is not always the top-ranking category, it would be interesting to see how *sf*-based weighting would perform in overlap rank-based fusion formulas.

For both WRS and OWRS formulas, three variations that amplify the contribution of the top performing system were investigated. These variations, in an increasing order of emphasis for the top system, are *Top System Pivot 1*(*pivot1*), *Top System Pivot 2*(*pivot2*), and *Overlap Boost*(*olpboost*). The basic idea here is to supplement the result of the best performing systems with fusion by using a weighting fusion function that amplifies the rank-based score of a document retrieved by the top systems while dampening the contributions from worse performing systems. A generalized form of *pivot1*, *pivot2*, and *olpboost* can be expressed as:

$$FS = \sum (wf_{kj}(L_i) * RS_i), \tag{11}$$

where:*FS* = fusion score of a document,

$w_{kj}(L_i)$ = weighting function of system group L_i in overlap partition *k* at rank *j,*

L_i= system group of system i based on performance

RS_i = rank-based score of a document by system i.

Equation 12, 13, and 14 describe the weighting functions of *pivot1*, *pivot2*, and *olpboost*. The equations essentially rerank the results of top systems only by setting the score of documents uniquely retrieved by non-top systems to zero. By using the interim fusion score that gets progressively larger with overlap, these formulas add more explicit emphasis on the overlap factor, which was leveraged only implicitly before in the summation process. *pivot2* adds more granularity in the weighting function to allow for the varying contribution levels of the overlapped systems, while *olpboost* adds yet another boost to top systems by multiplying their scores with the overlap count.

$$wf_{kj}(L_i)=\begin{bmatrix} avgp(i)*w_{ikj} & if & L_i = sg1 \\ fsc*w_{ikj} & if & L_i \neq sg1 \quad \& \quad olp1 \\ 0 & otherwise & \end{bmatrix}, \tag{12}$$

$$wf_{kj}(L_i)=\begin{bmatrix} avgp(i)*w_{ikj} & if & L_i = sg1 \\ fsc*w_{ikj} & if & L_i = sg2 \quad \& \quad olp1 \\ fsc*avgp(i)*w_{ikj} & if & L_i \neq (sg1 \| sg2) \quad \& \quad olp1 \\ 0 & otherwise & \end{bmatrix}, \tag{13}$$

$$wf_{kj}(L_i)=\begin{bmatrix} avgp(i)*w_{ikj}*ocnt & if & L_i = sg1 \\ fsc*w_{ikj} & if & L_i = sg2 \quad \& \quad olp1 \\ fsc*avgp(i)*w_{ikj} & if & L_i \neq (sg1 \| sg2) \quad \& \quad olp1 \\ 0 & otherwise & \end{bmatrix}, \tag{14}$$

where: $avgp(i)$ = overall average precision of system i in the training set,

$ocnt$ = the number of systems that retrieved a document

$sg1$ = the best systems

$sg2$ = second best performing systems

$olp1$ = true if document is retrieved by sg1

fsc = interim fusion score of a document[5)]

Ⅳ. Results

The study generated a massive amount of data that consisted of 66 single system results(36 VSM, 6 HITS, 24 TM) and numerous sets of fusion results. Since the focus of the study is on fusion methods, the discussion of results will center on fusion results preceded by a summary analysis of single system results.

1. Single System Results

The best performing single systems by average precision were: *vsmlb1.1* (VSM with long query, body text, phrase, and no feedback), *hitsmb1.1*(HITS with short host, seed set system of *vsmmb1.1*), and *tmlb1.1*(TM with top category, body text, phrase, no feedback). While rankings of top performing systems within each method remained consistent by various performance measures(Table 4), there were marked differences in performance across methods. In fact, the average precision values of top systems diminished roughly by half in each method order of VSM, TM, and HITS, thus indicating the overpowering advantage of text-based method over other methods. In general, the most influential system parameter appeared to be the query length. It is interesting to note that VSM and HITS systems benefit from longer queries, whereas TM systems

〈Tab. 4〉 Single System results

systems	*avgP*	*optF*	*R-P*	*P@5*	*P@10*	*P@20*	*P@100*	*P@200*
vsmlb1.1	.1529	.2528	.1959	.3500	.3150	.2520	.1306	.0902
vsmlb0.1	.1480	.2480	.1882	.3480	.3110	.2480	.1270	.0884
vsmlb1.2	.1473	.2485	.1895	.3300	.2960	.2420	.1266	.0888
hitsmb1.1	.0399	.1336	.0792	.0940	.1000	.0935	.0820	.0743
hitslb1.1	.0393	.1321	.0726	.0860	.0820	.0790	.0816	.0754
hitssb1.1	.0297	.1094	.0626	.0600	.0670	.0700	.0683	.0631
tm1b1.1	.0758	.1354	.1033	.1760	.1600	.1320	.0759	.0567
tm1b0.1	.0757	.1354	.1035	.1740	.1590	.1325	.0757	.0566
tm1b0.2	.0750	.1350	.1020	.1760	.1620	.1345	.0748	.0558

*avgP: average precision; optF: optimum F; R-P: R-Precision

5) When computing the fusion score, component scores are summed in system performance order(e.g. scores by *sg1* are added before *sg2*) to ensure consistant calculation of *fsc*.

perform better with shorter queries. Host definition, which determines the elimination of intrahost links and computation of link edge weights, seemed to be a crucial parameter for HITS systems.

1.1 Text-based Retrieval Results

Among the VSM system parameters tested, which are query length, term source, use of phrase terms, and use of pseudo-feedback, query length and term source were found to be most influential to the retrieval outcome. The influence of query length, which may be related to the amount of information, seems intuitive. Regardless of other parameter combinations, longer queries performed better than shorter queries in all cases except when system performances were degraded by the adverse effect of header text terms. The use of noun phrases, although helpful in 16 out of 18 system pairs, resulted in only a marginal increase in performance. Similarly, the use of pseudo-feedback resulted in a slight decrease in performance in most cases. Incidentally, the average precision of the best VSM system is roughly twice that of the top TM system, and four times the top HITS.

1.2 Link-based Retrieval Results

The results indicated the influence of host definition on retrieval performance for HITS systems. The shorter host definition is obviously far superior to longer definition(over 10 times better in average precision). As for the effect of the seed system on performance, one would expect better seed sets to produce better HITS results, since the quality of a seed set is amplified by link expansion much like the way the quality of “seed” documents is amplified by pseudo-feedback. The results, however, appear not to be totally consistent with such a supposition. The HITS tendency to degrade rather than enhance the seed system performance, even with the optimum seed set, may be due to incomplete relevance judgments and truncated link structure with heavy concentration of spurious links in the WT10g collection (Gurrinand Smeaton 2001).

1.3 Classification-based Retrieval Results

TM results showed only small performance variations across TM systems. TM parameter influences were quite orderly. All body text systems were ranked above body and header text systems. Within a given term source(e.g., body text, body and header text), systems using fewer

number of top categories were always ranked higher than systems using more categories. Within a given number of top categories and term sources, systems without pseudo-feedback always ranked above ones with feedback. Only the phrase use parameter show edin consistent results. Systems without phrase use ranked above ones with phrase use in five of the six body text system pairs, but only twice in six body and header text pairs.

2. Fusion Results

Since parametric combinations in each of VSM, HITS, and TM methods spawned a large number of systems(36 VSM, 6 HITS, and 24 TM systems), a brute force investigation of all possible fusion combinations was neither feasible nor desirable. Therefore, selective combinations of systems were conducted in a progressive fashion in an attempt to tease out the major factors that influence fusion.

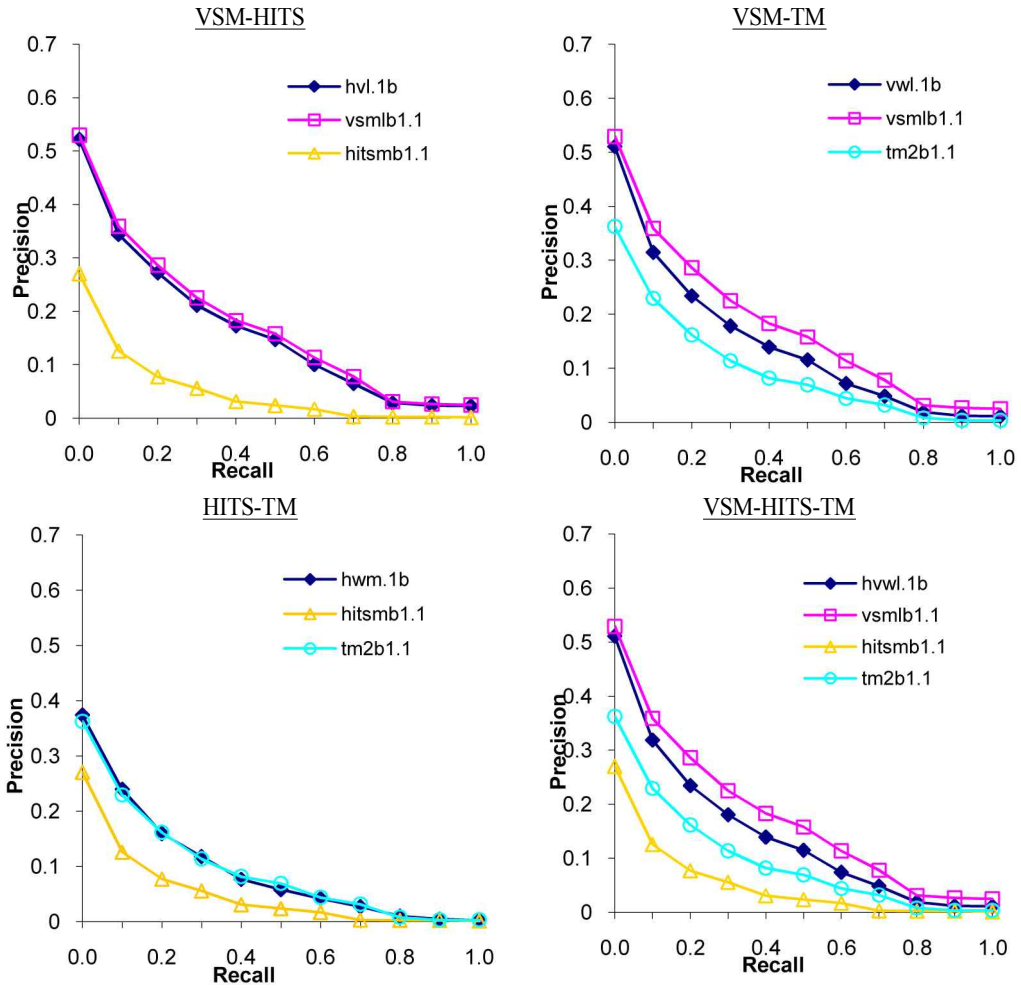
The results of the systems within each method(e.g., VSM, HITS, or TM) were combined first to see the effect of fusion without cross-method interplay of systems. Henceforth, combining systems within a given method will be referred to as *Intra-Method* fusion, as opposed to *Inter-Method* fusion that combines results across methods. In both intra- and inter-method fusion, systems were combined using *Weighted Rank Sum*(WRS) fusion formulas in a general fusion approach, which produced fusion component combinations of interest by focusing on system parameters that exhibited strong influences in single system runs. In contrast to the "every system for itself" approach of inter- and intra-method fusion, the third phase of fusion, called *Top System* fusion, took the "elitist" approach of combining only a handful of "best" systems from each method using variations of WRS formulas. The next three sections examine the results of intra-, inter-, and top system fusion.

2.1 Intra-Method Fusion

In both VSM and TM fusion, WRS fusion performance closely shadowed the baseline system. In HITS fusion, however, WRS surpassed the baseline performance at lower ranks. WRS and baseline results were almost identical in TM fusion. It is interesting to note that combining HITS results improved the retrieval performance, while little was gained by combining VSM or TM results. One possible explanation for this phenomenon may be that the combined solution space

of HITS systems is much larger than that of the best individual HITS system, while the best system dominates the combined solutions space in VSM and TM methods.

2.2 Inter-Method Fusion



<Fig.1> Recall-Precision Curve of Inter-Fusion Systems by Method Combination

In each of the four possible combinations of the three methods, fusion was conducted in a similar manner as the intra-method fusion to investigate the general fusion tendencies of cross-method fusion rather than to focus specifically on potentially advantageous system

combinations. Observations from intra-method fusion mostly held true in inter-method fusion, although fusion seemed to degrade the best single system performance more in inter-method fusion than in intra-method fusion. In all but the HITS-TM method combination, the baseline systems acted as upper and lower bound performance thresholds and fusion results fell nicely between them. There was, however, a distinct difference in the level of fusion results. As can be seen in Figure 1[6], which plots recall-precision curves of fusion systems with the highest average precision against the baseline system results, VSM-HITS fusion system results tended to be closer to the upper bound baseline while VSM-TM fusion results fell towards the middle. VSM-HITS-TM fusion results fell towards the middle of the upper and higher of the two lower bound baseline results.

Examination of overlap offers a possible explanation for this fusion outcome. High overlap counts for VSM and TM but low overlap count for HITS (Table 5) suggest that documents retrieved by VSM, which get boosted by the fusion formula, are much more likely to dominate the higher ranks of the VSM-HITS combined results than documents retrieved by the HITS systems. When VSM and TM system results are combined, however, documents retrieved by either system will get the overlap boost. Unfortunately, the performance level of VSM system may be degraded since documents of high ranks by the TM system are less likely to be relevant than those by VSM systems as implied by the lower performance levels of TM systems. In fact, the proportion and the size of non-relevant documents with high overlap were much larger in TM than in VSM systems, which may very well account for the adverse effect of TM systems on fusion results.

<Tab. 5> Overlap Statistics for all systems at rank 100

olp	olpV	olpH	olpW	%rel	olp	olpV	olpH	olpW	%rel	olp	olpV	olpH	olpW	%rel
65	36	5	24	33	45	22	2	20	19	25	13	1	11	6
64	36	4	24	14	44	26	0	17	18	24	16	0	6	6
63	36	3	24	21	43	23	2	17	18	23	12	1	9	6
62	36	2	24	30	42	22	1	17	18	22	14	1	6	7
61	34	2	24	24	41	21	1	17	28	21	14	1	5	3
60	35	0	23	27	40	23	0	15	13	20	14	0	5	2

6) System name prefixes indicate the methods combined (i.e. "*v*" for VSM, "*h*" for hits, "*w*" for TM)

OLP = number of systems that retrieved a document
OLPV = number of VSM systems that retrieved a document
OLPH = number of HITS systems that retrieved a document
OLPW = number of WD (TM) systems that retrieved a document
%REL = percentage of relevant document in a partition defined by OLP

2.3 Top System Fusion

While intra- and inter-fusion experiments study the general effect of comprehensive fusion, *Top System Fusion* explores the effect of combining a handful of "top" systems from each method with various fusion formulas. Limiting fusion component systems to a small set of high-performance systems reduces the number of system interactions while capturing the most significant system contributions to the fusion process, thus creating an environment more suited for the fine tuning of the fusion formula than the massive fusion approach of intra- and inter-method fusions.

The fusion formulas tested in top system fusion are enumerated below. In the first column are the suffixes attached to system names to indicate the fusion formula used.

a	WRS
ba	OWRS, *no pivot*
bb	OWRS, *pivot1*
bc	OWRS, *pivot2*
bd	OWRS, *olpboost*
ca	ROWRS-*sf*, *no pivot*
cb	ROWRS-*sf*, *pivot1*
cc	ROWRS-*sf*, *pivot2*
cd	ROWRS-*sf*, *olpboost*
da	ROWRS-*F*, *no pivot*
ea	ROWRS-*P*, *no pivot*

Three different rank-based system performance measures were tested for the ROWRS formula. ROWRS-*sf* used the Success/Failure measure based on the retrieval success/failure at each rank to compute the fusion component weights, while ROWRS-*F* used the Effectiveness measure and ROWRS-*P* used precision. For the OWRS and ROWRS-*sf* formulas, the effects of the "top system pivot[7])" was investigated by testing *pivot1* (Equation 12), *pivot2* (Equation 13), and

olpboost (Equation 14) variations against the baseline (*no pivot*).

Table 6, which lists the best results from each fusion formula along with the baseline of the best single system result[8] in a descending order of average precision, shows that fusion formulas ROWRS-*sf* with *olpboost*, ROWRS-*sf* with *pivot1*, and ROWRS-*sf* with *pivot2*, in that order outperform the best single system results. The gain in performance, however, is marginal as fusion results are tightly grouped around the baseline single system result. All top three fusion systems retrieved fewer relevant documents and had higher precision at 200 than the baseline, which suggests that the gain in performance came from boosting the ranking of relevant documents at earlier ranks. The loss in the number of relevant documents retrieved can be attributed to ROWRS formula's tendency to eliminate uniquely retrieved documents from the result set. Even without the uniquely retrieved relevant documents, ROWRS outperforms OWRS regardless of top-system pivot variations.

〈Tab. 6〉 Top System Fusion results

systems	*avgP*	*optF*	*R-P*	*P@5*	*P@10*	*P@20*	*P@100*	*P@200*
Fhsl1F2cd	.1739	.2679	.2016	.3240	.2980	.2350	.1108	.0743
F2hsl1c31cb	.1721	.2654	.2076	.3560	.2960	.2260	.1084	.0718
F2hsl1c31cc	.1721	.2654	.2076	.3560	.2960	.2260	.1084	.0718
vsmlb1.1	.1652	.2592	.1969	.3280	.2980	.2280	.1064	.0710
F2hsl1c31ca	.1635	.2571	.2039	.3400	.3020	.2080	.1068	.0735
F2hsl1c31ba	.1635	.2561	.2025	.3320	.2820	.2190	.1040	.0697
Fhsl1c31bb	.1613	.2667	.1984	.3000	.2600	.2150	.1074	.0749
Fhsl1c31bc	.1613	.2667	.1984	.3000	.2600	.2150	.1074	.0749
Fhsl1c31bd	.1613	.2665	.1983	.3000	.2580	.2150	.1074	.0746
ql1hsl1c31da	.1581	.2554	.1891	.3440	.2760	.2200	.1026	.0690
F2hsl1c21a	.1578	.2515	.1931	.3120	.2660	.2220	.1076	.0732
ql1hsl1c31ea	.1564	.2524	.1855	.3400	.2940	.2160	.1034	.0691

Comparison of rank-based measures shows the success/failure measure to be superior to precision- or effectiveness-based measures for ROWRS. As for top-system pivot variations, the

7) "Top system pivot" refers to *pivot1*, *pivot2*, and *olpboost* described in section 3.2.4.

8) "Best" or "top" implies best or top performance, which is normally measured by average precision unless explicitly stated otherwise.

ROWRS formula seems to work best with the heaviest emphasis on the top system contribution (*olpboost*), in contrast with the OWRS formula that shows the best results without any top-system emphasis(*no pivot*).The different effects of top-system pivot between OWRS and ROWRS formulas may indicate the relationship between the rank and the relevance of a document in top systems.

3. Overlap Analysis

In section 4.2.1 where intra-fusion results were discussed, it was suggested that HITS systems had the most to gain by fusion due to their diverse solution spaces. One way to confirm such a hypothesis is to examine the degree of overlap in relevant documents retrieved by HITS systems. Table 7lists the total number of relevant documents retrieved(RRN) as well as the number of relevant documents uniquely retrieved by a system within a given method(e.g. VSM, HITS, TM), and thus describe the degree of overlap in relevant documents retrieved. The VSM, HITS, and TM columns indicate that the solution spaces for HITS systems overlap much less than those of VSM or TM systems. More specifically, the unique contributions of the top 3 HITS systems, which are considerably larger than those of the top 3 VSM or TM systems, imply that the HITS method has the most to gain by fusion.Examination of the inter-method fusion results in section 4.2.2 also merited overlap analysis. Larger numbers in the H-W column of Table 7 indicate the greater potential gain for HITS-TM fusion, which we saw in section 4.2.2.

<Tab. 7> Number of Relevant Documents Retrieved at rank 1000

System	*RRN*	*VSM*	*HITS*	*WD*	*V-H*	*V-W*	*H-W*
vsmmb1.1	3303	1	-	-	1	1	-
vsmmb0.1	3255	3	-	-	2	2	-
vsmmb1.2	3247	3	-	-	3	3	-
tm1b0.1	2243	-	-	0	-	0	0
tm1b1.1	2240	-	-	0	-	0	0
tm1b0.2	2231	-	-	3	-	1	2
hitslb1.1	1886	-	247	-	9	-	94
hitsmb1.1	1775	-	72	-	3	-	26
hitssb1.1	1598	-	119	-	7	-	25

<Tab. 8> Optimum Performance Levels

At rank 1000			*At rank 20*		
System	*avgP*	*RRN*	*System*	*avgP*	*RRN*
Fall	0.7819	1725	Fall	0.3361	492
Fvw	0.7769	1710	Fvw	0.3212	469
Fvh	0.7618	1629	Fvh	0.2979	402
Fvsm	0.7555	1610	Fvsm	0.2828	378
Fhw	0.6716	1397	vsmlb1.1	0.2100	228
vsmmb1.1	0.6398	1340	Fhw	0.1732	322
Fwd	0.5719	1206	Fwd	0.1162	198
Fhits	0.5084	914	Fhits	0.0720	156
tm1b0.1	0.4741	948	tm1b0.2	0.0825	133
hitslb1.1	0.4381	724	hitsmb1.1	0.0391	87

The fact that VSM-HITS fusion results were closer to the upper bound(defined by the best VSM system) while VSM-TM fusion results fell more towards the middle of the upper and lower bound(defined by the best VSM and TM systems) requires different kind of overlap analysis to explain. The overlap statistics table(Table 5) shows fairly even number of overlap counts for VSM and TM(OLPV and OLPW columns) but hardly any overlap count for HITS(OLPH column) at high ranks. Even at lower ranks, documents retrieved by many HITS systems are small in numbers compared to VSM and TM. Since fusion formulas reward the overlapped documents, documents retrieved by VSM systems are much more likely to influence the VSM-HITS combined results than documents retrieved by HITS systems. Thus, the VSM-HITS fusion results are closer to the VSM baseline than the HITS baseline. When VSM and TM system results are combined, however, documents retrieved by either system get the overlap boost and the results of VSM systems get degraded by the large number of non-relevant documents with high overlap in TM systems.

Table 8, which describes the maximum potential for fusion, shows that combining all VSM systems(*Fvsm*) could increase the optimum average precision of the best VSM system from 0.6398 to 0.7555 by introducing 270 more relevant documents to the solution space. Combining all systems of all methods further raise the maximum fusion potential to the average precision of 0.7819 with 1725 total relevant documents retrieved. In other words, the numbers in optimum performance level tables proves the existence of the fusion potential by showing that combining the retrieval results of individual systems “can” increase the total number of relevant documents retrieved.

Table 9, which summarizes the overlap statistics tables to display the density of relevant documents at various ranks and overlap, show the higher relevance density(i.e. proportion of relevant documents in a given overlap) for not only higher overlap but also higher ranks. Unfortunately, the relevance density is below 50%, which suggests that overlap without the consideration of document ranking and systems that retrieved the overlapped document may not always be a good indicator of relevance. In fact, more documents are apt to be non-relevant than relevant for a given overlap in the current experiment, although more overlapped documents are more likely to be relevant than less overlapped documents.

Table 10, which relates overlap with not only relevance but also document ranks, shows that in general, non-relevant documents are ranked lower than relevant documents with the same overlap in VSM and TM systems but the reverse is true for HITS systems. This peculiar pattern of overlap in HITS systems may explain why the rank-based fusion formula did not do well in HITS fusion.

〈Tab. 9〉 Relevance Density in Overlapped Documents for all systems, Topics 451−500

	overlap >=10		*overlap >=20*		*overlap >=30*		*overlap >=40*		*overlap >=50*	
rank	*N**	*relp***	*N*	*relp*	*N*	*relp*	*N*	*relp*	*N*	*relp*
5	429	0.22	172	0.32	77	0.38	41	0.44	13	0.62
10	913	0.19	355	0.28	164	0.33	94	0.38	36	0.44
20	1902	0.13	740	0.22	345	0.28	205	0.35	74	0.43
100	10384	0.06	4423	0.10	2161	0.15	1247	0.18	560	0.20
200	20047	0.04	9315	0.07	4829	0.10	2707	0.13	1485	0.15
1000	92608	0.02	47115	0.03	26203	0.04	14657	0.05	6444	0.07

*N = total number of documents retrieved by 10 or more systems
**relp = proportion of relevant documents in N documents.

〈Tab. 10〉 Average Ranks in Overlapped Documents for all systems with Overlap >= 10*

rank	*Topics 451-500*									*Topics 501-550*								
	N	*p*	*pV*	*pH*	*pW*	*avg R*	*avg RV*	*avg RH*	*avg RW*	*N*	*p*	*pV*	*pH*	*pW*	*avg R*	*avg RV*	*avg RH*	*avg RW*
5	429	.4	.3	.2	.6	3	2	.2	1	460	.4	.1	.3	.6	3	3	0.1	1.6
10	913	.3	.5	.36	.6	6	5	.4	2	947	.3	.2	.4	.5	6	4	0.4	2.5
20	1902	.4	.4	.51	.6	12	10	1	6	1958	.6	.3	.3	.5	12	10	0.4	5.7
100	10384	.5	.7	.36	.7	59	51	11	20	10516	.8	.5	.3	.7	60	51	13.4	20.8
200	20047	.7	.6	.23	.7	120	101	22	44	20984	.4	.3	.1	.4	122	107	33.5	42.3
1000	92608	.7	.6	.51	.7	625	530	33	238	98443	.8	.7	.4	.8	610	512	34.5	211.7

* Column *p* (*pV, pH, pW*) shows proportion of non-relevant documents whose average ranks (of VSM, HITS, WD systems) are larger than that of relevant documents with the same overlap.

Ⅴ. Concluding Remarks

In this study, we explored the question of whether combining text-, link- and classification-based retrieval methods can improve the Web search performance by examining the effects of combining retrieval results of text-, link-, and classification-based retrieval systems using the WT10g test collection and Yahoo directory information. The retrieval results of text-based systems based on the Vector Space Model, link-based systems using the HITS algorithm, and classification-based systems using Yahoo categories were combined using a rank-based fusion formula. In addition, a handful of the best performing systems from each method were combined with variations of the rank-based fusion formulas to explore the optimization of fusion parameters.

The differences in retrieval methods that affected different retrieval outcomes appeared to influence both intra- and inter-method fusion, where the system results were combined within and across retrieval methods. Interestingly, the only intra-method fusion that enhanced the baseline performance of the best fusion component results occurred with the worse performing HITS systems. Intra-method fusion of VSM and TM systems behaved similarly in that fusion detracted from the baseline performance although combining TM system results degraded baseline results much more severely than VSM fusion when using the SM formula.

To investigate the possible reasons why combining HITS system results enhanced retrieval performance while combining VSM or TM system results degraded the baseline performance, we examined the degree of overlap in relevant documents in HITS systems in comparisons with VSM and TM systems and found that HITS systems retrieved much more diverse sets of relevant documents than VSM or TM systems and thus had the most to gain by fusion.

In top system fusion, where variations of WRS formulas were used to combine the results of a few top systems from each method in an attempt to improve the retrieval performance of the top system results while minimizing both the computational overhead and the adverse contributions from the poor performing systems, the fusion succeeded in enhancing the best single system result. Top system pivot with the overlap boost, which emphasized progressively the contributions of overlapped top system documents showed the best results, which suggests that leveraging overlap in conjunction with the rankings of the best performing systems is an advantageous fusion approach.

Perhaps the most significant findings of the study came from the overlap analysis, which revealed that the total number of relevant documents in the combined result sets of VSM, HITS, and TM systems were much more than the largest number of relevant documents retrieved by any single system. This observation disputes the implicit null hypothesis against fusion(i.e., there is nothing to be gained by combining single system results) and thus strongly suggests that the solution spaces of text-, link-, and classification-based retrieval methods are diverse enough for fusion to be beneficial. It is important to note that HITS runs, despite their lower performance levels than VSM and TM runs, appeared to have the most unique contributions to the fusion pool. The high degree of unique contributions by HITS systems could be a reflection of its retrieval approach, which is distinct from VSM and TM systems with heavy reliance on text-based retrieval techniques.

One of the most important issues for fusion is the optimization of the fusion formula. Given less than the optimum results by individual systems, how can we combine them to bring up the ranking of the relevant documents? We have seen in the overlap analysis that, although the documents retrieved by more systems are more likely to be relevant, the simple number of systems that retrieve a document is not a good indicator for relevance since highly overlapped documents were often more likely to be non-relevant than relevant. One way to compensate for this is to rely on top performing systems as was done in top system fusion. Top system fusion, however, tends to ignore the unique contributions with its heavy emphasis on overlap. One of the most difficult challenges of top system fusion, as well as fusion in general, is devising a method that rewards both the overlapped and unique contributions to the combined solution space.

The study selected retrieval methods that leverage three distinct sources of evidence on the Web and implemented a variety of ordinary systems and combined their retrieval results using ad-hoc variations of common fusion formulas. Although the performance of the best fusion result was only marginally better than the best individual result, the analysis of overlap strongly suggested that the solution spaces of text-, link-, and classification-based retrieval methods were diverse enough for fusion to be beneficial. Furthermore, analysis of the results revealed much insight concerning the effects of system parameters, the relationship between overlap, document ranking and relevance, and important characteristics of the fusion environment. In addition to establishing the existence of the fusion potential for Web IR, this study has provided a rich foundation on which to continue the exploration of fusion in future research.

There are almost as many possibilities for future research in fusion as there are fusion combinations and fusion formulas. The main contributions of this study are twofold. First, it confirms the viability of fusion for Web IR by not only determining the existence of the fusion potential in the combined solution spaces of text-, link-, and classification-based retrieval methods but also by demonstrating that relatively simple implementation of fusion does improve the retrieval performance. Furthermore, the study lays the groundwork for future research, where the current research framework can be extended in many dimensions to explore various aspects of fusion.

References

Bartell, Brian T., G. W. Cottrell and R. K. Belew. 1994. "Automatic combination of multiple ranked retrieval systems." *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.*

Belkin, Nicholas J., C. Cool, W. B. Croft and J. P. Callan. 1993. "The effect of multiple query representations on information retrieval system performance." *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval,* 339-346.

Bharat, Krishnaand M. R. Henzinger. 1998. "Improved Algorithms for Topic Distillation in Hyperlinked Environments." *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval,* 104-111.

Brin, Serge andL. Page. 1998. "The anatomy of a large-scale hyper textual Web search engine." *Computer networks and ISDN systems,* 30(1): 107-117.

Buckley, Chris, G. Salton, J. Allan and A. Singhal. 1995. "Automatic query expansion using SMART: TREC 3." In D. K. Harman (Ed.), *The Third Text Rerieval Conference (TREC-3)* (NIST Spec. Publ. 500-225, pp.1-19). Washington, DC: U.S. Government Printing Office.

Buckley, Chris, A. Singhal and M. Mitra. 1997. "Using query zoning and correlation within SMART: TREC 5." In E. M. Voorhees & D. K. Harman (Eds.),

The Fifth Text REtrieval Conference (TREC-5) (NIST Spec. Publ. 500-238, pp. 105-118). Washington, DC: U.S. Government Printing Office.

Buckley, Chris, A. Singhal, M. Mitra and G. Salton. 1996. "New retrieval approaches using SMART: TREC 4." In D. K. Harman (Ed.), *The Fourth Text REtrieval Conference (TREC-4)* (NIST Spec. Publ. 500-236, pp. 25-48). Washington, DC: U.S. Government Printing Office.

Chakrabarti, Soumen, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson and J. Kleinberg. 1998. "Automatic resource list compilation by analyzing hyperlink structure and associated text." *Proceedings of the 7th International World Wide Web Conference.*

Fishburn, Peter C. 1970. *Utility theory for decision making.* New York: John Wiley & Sons.

Fox, Edward A. andJ. A. Shaw. 1994. "Combination of multiple searches." In D. K. Harman (Ed.), *The Second Text Rerieval Conference (TREC-2)* (NIST Spec. Publ. 500-215, pp.243-252). Washington, DC: U.S. Government Printing Office.

Fox, Edward A. and J. A. Shaw. 1995. "Combination of multiple searches." In D. K. Harman (Ed.), *The Third Text Rerieval Conference (TREC-3)* (NIST Spec. Publ. 500-225, pp. 105-108). Washington, DC: U.S. Government Printing Office.

Frakes, Williams B. and R.Baeza-Yates.eds. 1992. *Information retrieval: Data structures & algorithms.* Englewood Cliffs, NJ: Prentice Hall.

Gurrin, Cathal and A. F.Smeaton. 2001. "Dublin City University experiments in connectivity analysis for TREC-9." In E. M. Voorhees & D. K. Harman (Eds.), *TheNineth Text Rerieval Conference(TREC-9).* Washington, DC: U.S. Government Printing Office.

Katzer, Jeffrey, M. J. McGill, J. A. Tessier, W. Frakes and P. DasGupta. 1982. "A study of the overlap among document representations." *Information Technology: Research and Development, 1,* 261-274.

Keen, E. Michael. 1973. “The Aberystwyth index languages test.” *Journal of Documentation, 29,* 1-35.

Kleinberg, Jon. 1999. “Authoritative sources in a hyperlinked environment.” *Journal of the Association for Computing Machinery, 46(5),* 604-632.

Lee, Joon Ho. 1996. *“Combining multiple evidence from different relevance feedback methods(Tech. Rep. No.IR-87).”* Amherst: University of Massachusetts, Center for Intelligent Information Retrieval.

Lee, Joon Ho. 1997. “Analyses of multiple evidence combination.” *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval,* 267-276.

Modha, Dharmendra and W. S. Spangler. 2000. “Clustering hypertext with applications to Web searching.” *Proceedings of the 11th ACM Hypertext Conference,* 143-152.

Page, Larry, S. Brin, R. Motwani and T. Winograd.1998. “The Page Rank citation ranking: Bringing order to the Web.” *Technical Report,* Stanford Digital Library Technologies Project.

Plaunt, Christian and B. A. Norgard. 1998. “An Association Based Method for Automatic Indexing with a Controlled Vocabulary.” *Journal of the American Society for Information Science,* 49(10): 888-902.

Saracevic, Tefko and P. Kantor. 1988. “A study of information seeking and retrieving. III. Searchers, searches, overlap.” *Journal of American Society for Information Science,* 39: 197-216.

Singhal, Amit, C. Buckley and M. Mitra. 1996. “Pivoted document length normalization.” *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval,* 21-29.

Smith, Linda. C. 1979. *Selected Artificial Intelligence Techniques in Information Retrieval Systems Research.* Ph. D. diss., Syracuse University, U. S.

Sparck Jones, Karen. 1974. “Automatic indexing.” *Journal of Documentation 30,* 393−432.

Sumner, Robert. G., K. Yang, R. Akers and W. M. Shaw. 1998. “Interactive retrieval using IRIS: TREC-6 experiments.” In E. M. Voorhees & D. K. Harman(Eds.), *The Sixth Text REtrieval Conference(TREC-6).*

Vogt, Christopher. C and G. W. Cottrell. 1998. “Predicting the performance of linearly combined IR systems.” *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval,* 190-196.

Williams, Martha E. 1977. “Analysis of terminology in various CAS data files as access points for retrieval.” *Journal of Chemical Information and Computer Sciences,* 17: 16-20.

Wong, S. K. Michael, Y. Y. Yao and P.Bollmann. 1988. “Linear structure in information retrieval.” *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 219-232.

Wong, S. K. Michael, Y. Y. Yao, G. Salton and C. Buckley. 1991. “Evaluation of an adaptive linear model.” *Journal of the American Society for Information Science,* 42: 723-730.

Yang, Kiduk. 2005. “Information retrieval on the web.” *ARIST,* 39(1): 33-80.