Surrogate Scoring for Improved Metasearch Precision

Steven M. Beitzel, Eric C. Jensen, Ophir Frieder
Information Retrieval Laboratory
Illinois Institute of Technology
{steve,ej,ophir@ir.iit.edu}

Abdur Chowdhury, Greg Pass
Search & Navigation
America Online
{cabdur, gregpass1@aol.com}

ABSTRACT

We describe a method for improving the precision of metasearch results based upon scoring the visual features of documents' surrogate representations. These surrogate scores are used during fusion in place of the original scores or ranks provided by the underlying search engines. Visual features are extracted from typical search result surrogate information, such as title, snippet, URL, and rank. This approach specifically avoids the use of search engine-specific scores and collection statistics that are required by most traditional fusion strategies. This restriction correctly reflects the use of metasearch in practice, in which knowledge of the underlying search engines' strategies cannot be assumed. We evaluate our approach using a precision-oriented test collection of manually-constructed binary relevance judgments for the top ten results from ten web search engines over 896 queries. We show that our visual fusion approach significantly outperforms the rCombMNZ fusion algorithm by 5.71%, with 99% confidence, and the best individual web search engine by 10.9%, with 99% confidence.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms

Algorithms, Performance, Experimentation.

Keywords

Data Fusion, Machine Learning, Metasearch, Web Search

1. INTRODUCTION

The goal of metasearch is to retrieve search results from multiple component search engines and construct a single, high-quality, unified ranked list. In practice, component search engines rarely return relevance scores or collection statistics in their results; therefore, the only information available as a basis for unifying the results from these engines is the limited document representation shown to the user, called the surrogate. Surrogates are designed to contain enough information for a human searcher to determine whether or not the represented document is relevant to the search in question. The surrogates for most web search engines include a title, a URL, a short snippet (keyword in context), and the rank. None of the ten engines used in this study supplied relevance scores.

The goal of this study is to develop a method of "visual fusion" that uses the visual cues present in surrogates as the sole source of information for fusing component engine results into a single ranked list. In essence, the surrogate itself is scored for perceived relevance, and that surrogate score is used as the evidence for the fusion algorithm. This method enables metasearch systems to perform fusion without requiring any knowledge of the underlying models of retrieval used by each component engine, as opposed to previously-studied fusion algorithms, such as CombMNZ [1], that require underlying relevance scores. The strategy of our approach is to identify as many visual features as possible and determine the relative advantage of each feature using a machine learning algorithm. We hope this model correlates with the decision-making process of a human user visually evaluating surrogates for relevance.

2. PRIOR WORK

Data fusion in IR has been studied for several years; a survey and analysis of prior fusion research can be found in [2]. We know of no work that is specifically focused on using visual features of surrogates. Many commonly accepted fusion algorithms such as CombMNZ and its variants are not applicable to this problem because they rely on the presence of relevance scores assigned to each document by the underlying engine.

Some methods have been developed that rely only on search result rank, such as rCombMNZ and Condorcet-fuse [3]. These methods were evaluated using TREC-style collections with very deep document pools and a large number of component engines, unlike the web-oriented test collection used in our study. It is clear that more sophisticated features could be extracted if the fusion system were to crawl and analyze the full text of each result document; however, this is likely to exceed the operational requirements of a typical web metasearch system [4].

Additionally, many commonly-used fusion methods are unsupervised, relying solely on the scores or ranks from the underlying search engines to provide an estimate of the relevance of a given document. Some supervised versions of these algorithms do exist (e.g., WeightedCondorcet-fuse, WeightedCombMNZ), but the extent of their supervision is typically restricted to a linear combination in which each system's confidence in a document (be it rank or score) is weighted by that system's performance over some training set, as judged by an arbitrary metric (such as mean average precision).

3. METHODOLOGY

We propose a method of fusion that uses the visual features of surrogates and supervised machine learning to produce a fused list of results. An overview of the methodology is as follows:

1. Construct a training set consisting of search results over a number of queries, and manually evaluate the documents for binary relevance.
2. For each document in the training set:
   a. Extract the set of visual features from the document surrogate.
   b. Use the visual features and binary relevance judgment as a training instance for a supervised machine learning algorithm.
3. Execute the learning algorithm to determine the learned weight for each visual feature.

SIGIR 2005, August 15--19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

4. Construct a testing set consisting of search results over a number of queries.
5. For each document in the testing set:
   a. Calculate the document's surrogate score by summing the products of the visual feature weights and values over all surrogate features.
   b. Insert the document into the fused list of results, sorted by its surrogate score.
6. Evaluate the final fused list of results using applicable measures.

4. EXPERIMENTATION & RESULTS

To appropriately evaluate a task designed for web metasearch, we used an available precision-oriented test collection designed to closely model the web environment [5]. The collection consists of the top ten search results from ten web search services, manually evaluated for binary relevance over 896 random queries submitted to AOL™ Search. The ten web search engines were Google™, Yahoo™, Wisenut™, Teoma™, Altavista™, AllTheWeb™, Lycos™, Gigablast™, MSN™, and the MSN™ TechPreview (hereafter referred to, in no particular order, as E1-10). We designated 2/3 of this collection as the training set and 1/3 as the testing set.

Over the entire collection, the average number of unique results for a given query across all ten engines was 43, and the mean number of results in common between any two engines was 37%. This figure gives some indication of the maximum impact a fusion algorithm based on multiple evidence, such as rCombMNZ, can have.

We extracted 31 visual features from the surrogate representation of each result document and learned their relative weights using the unmodified SMO [6] machine learning implementation in WEKA (http://www.cs.waikato.ac.nz/ml/weka/) with a decision threshold constant of 1.310. The following list orders the 31 features by weight, from the most positive indication of relevance (positive numbers) to the most positive indication of non-relevance (negative numbers): % of query character ngrams in the title (2.309), average distance between query terms in the title (1.064), % of title character ngrams in the query (0.555), % of query terms in the title (0.397), URL path depth (0.188), % of query character ngrams in the snippet (0.128), % of query character ngrams in the URL (0.059), E1 (0.035), E6 (0.029), E2 (0.025), % of snippet term ngrams in the query (0.025), % of query terms in the snippet (0.019), number of terms in the query (0.017), % of snippet terms in the query (0.011), % of query term ngrams in the snippet (0.007), E4 (0.001), E5 (-0.001), E3 (-0.001), E10 (-0.008), URL contains query (-0.009), number of terms in the snippet (-0.011), E8 (-0.016), E9 (-0.028), E7 (-0.038), average distance between query terms in the snippet (-0.039), % of snippet character ngrams in the query (-0.055), % of title terms in the query (-0.080), number of terms in the title (-0.102), original rank (-0.151), % of title term ngrams in the query (-0.452), and % of query term ngrams in the title (-0.595).

We observed that the four most positive indicators of relevance were all title-based visual features; we also observed that the search engine itself, in most cases, was not a significant indicator of relevance.

We evaluated our fused list of results (vCombMNZ), the list of results produced by rCombMNZ, and each individual web search engine using mean average precision and precision@10. These results, given in Table 1, show that our fusion technique outperforms rCombMNZ by a relative improvement of 5.71% and the best of the individual engines by nearly 11% when using mean average precision. Furthermore, we calculated the error rate [7] of mean average precision over 2,401 sub-samples of 250 queries from the testing set (sampling error of 6.1% with 95% confidence), and found that the ranking of engines produced by mean average precision has an error rate of only 3.8%, with a 4.5% tie-rate at an absolute fuzziness of 0.3% difference in engines' mean scores. We also determined that the difference between visual fusion and both rCombMNZ and the best individual engine is statistically significant with 99% confidence over the original test set and all 2,401 sub-samples when using the non-parametric Wilcoxon paired signed-rank test.

Table 1: Mean Average Precision and P@10 for Visual Fusion, rCombMNZ, and Individual Engines

	P@10	MAP@10	Imp over best Ex MAP	Imp over avg Ex MAP
vCombMNZ	**0.742**	**0.693**	**10.9%**	**18.9%**
rCombMNZ	*0.704*	*0.655*	*4.9%*	*12.4%*
E1	*0.685*	*0.625*	*0.0%*	*7.2%*
E2	0.668	0.607	-2.9%	4.1%
E3	0.666	0.599	-4.1%	2.8%
E4	0.666	0.598	-4.3%	2.6%
E5	0.661	0.592	-5.2%	1.6%
E6	0.652	0.577	-7.6%	-0.9%
E7	0.625	0.571	-8.6%	-2.1%
E8	0.628	0.567	-9.2%	-2.7%
E9	0.629	0.555	-11.2%	-4.8%
E10	0.613	0.537	-14.0%	-7.8%

5. CONCLUSIONS

This paper describes a method of fusing results from multiple search engines that relies only on visual features of the documents' surrogate representations, and not upon knowledge of the underlying engines' models of retrieval and ranking. This approach is increasingly germane as the scale of metasearch grows to federate hundreds or thousands of content sources operating under disparate ranking methodologies. Surrogate scoring is an effective step forward in addressing this developing concern.

6. REFERENCES

[1] Fox, E. and Shaw, J.A. *Combination of Multiple Searches*. NIST, TREC-2, 1994.

[2] Beitzel, S., et. al. *Fusion of Effective Retrieval Strategies in the Same Information Retrieval System.* JASIST, 2004.

[3] Montague, M. and J. Aslam. *Condorcet Fusion for Improved Retrieval*. ACM-CIKM, 2002.

[4] Chowdhury, A. and G. Pass. *Operational Requirements for Scalable Search Systems*. ACM-CIKM, 2003.

[5] Jensen, E., et. al. *A Framework for Determining Necessary Query Set Sizes to Evaluate Web Search Effectiveness.* WWW, 2005.

[6] J. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization.* Advances in Kernel Methods, Support Vector Learning. MIT Press, 1999.

[7] Buckley, C. and E. Voorhees. *Evaluating Evaluation Measure Stability*. ACM-SIGIR, 2000.