Effects of Highly Agreed Documents in Relevancy Prediction *

Andrés R. Masegosa
Department of Computer Science and A.I., University of Granada, Spain.
andrew@decsai.ugr.es

Hideo Joho
Department of Computing Science, University of Glasgow, UK.
hideo@dcs.gla.ac.uk

Joemon M. Jose
Department of Computing Science, University of Glasgow, UK.
jj@dcs.gla.ac.uk

ABSTRACT

Finding significant contextual features is a challenging task in the development of interactive information retrieval (IR) systems. This paper investigated a simple method to facilitate such a task by looking at aggregated relevance judgements of retrieved documents. Our study suggested that the agreement on relevance judgements can indicate the effectiveness of retrieved documents as the source of significant features. The effect of *highly agreed documents* gives us practical implication for the design of adaptive search models in interactive IR systems.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Measurement, Experimentation

Keywords: Relevance prediction, highly agreed documents

1. INTRODUCTION

People disagree on the judgement of document relevancy [9]. However, the judgement of highly relevant documents are more likely to be agreed than that of partially relevant documents [8]. Therefore, when multiple judgements are available for document relevancy, the degree of relevance is likely to be indicated by the level of agreement on judgements. In other words, highly agreed documents can be seen as highly (non-)relevant documents.

In this paper, we hypothesise that highly agreed documents can facilitate the mining of significant contextual features. We defined a contextual feature as a variable that increased an information retrieval (IR) system's power of discriminating relevant documents from non-relevant ones. Therefore, one can measure the effect of contextual features based on the accuracy of document relevancy prediction. Finding significant contextual features has several implications for the design of effective IR systems. In particular, we aimed to contribute to methodological advance for the development of context-aware interactive IR systems [5]. An effective way to elicit significant features from a wide range of potentially relevant factors can help us make an IR system adaptive to a search environment. This paper investigates an approach for facilitating the process of finding significant features based on aggregated relevance judgements made by searchers. The rest of the paper is structured as follows. Section 2 discusses our methodology to vary the level of agreement on relevance judgements to test our hypothesis. Section 3 presents the results of our experiment and discusses the implications of highly agreed documents for the design of adaptive search models.

*This work was supported by ALGRA project (TIN2004-06204-C03-02), FPU scholarship (AP2004-4678), and EPSRC (EP/C004108/1).

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

Table 1: Click-through (CT) data

	No. of Docs	CT	Non-Rel (%)	Rel (%)
CT=1	605	605	46.6	53.4
CT=2	84	184	40.8	59.2
CT>2	48	256	46.5	53.5
All CT	737	1045	49.5	50.5

2. METHODOLOGY

Our overall approach was to use machine learning techniques as a diagnostic tool to measure the effect of highly agreed documents in relevancy prediction. In the experiment, four well-known probabilistic classifiers [2, 10, 4, 7] were used to predict document relevancy. Unlike the work in [3, 1], we used multiple classifiers since a single classifier was unlikely to show the significance of potential features in a complex dependency structure. Our evaluation was based on experimental data collected in a laboratory-based user study with 24 participants searching for four different topics independently [6]. In each topic, they were given up to 15 minutes to complete a search session. Participants were asked to bookmark a document when perceived relevant information was found. Both the documents which participants visited from search results (i.e., click-through documents) and the bookmarked (BM) ones were used to form varied levels of agreement on document relevancy.

The distribution of click-through (CT) data is shown in Table 1. As can be seen, a total of 1045 click-through actions were recorded on 737 unique documents. Of those, 58% of click-through were recorded on the documents which had a single click-through (CT=1). While the portion of relevance judgements varied over the frequency of click-through, the overall performance was approximately 50%. From the in-

Table 2: Categorisation of candidate features.

Category	Example	Size
Object features		116
Document Textual Features	No. of words	13
Visual Appearance	No. of CSS links	16
Visual HTML tags/att. tags	No. of bold tags	17
Layout Features	No. of tables	14
Structural Features	URL domain	10
Selective Words	Word 'help'	22
Special HTML tags/att. tags	No. of meta tags	24
Interaction Features	Query Length	5

Table 3: Relevance aggregation method (NA: Negative agreement, PA: Positive agreement)

Condition	Non Relevant	Relevant	Discarded
C_1	NA > 50%	PA > 50%	Otherwise
C_2	NA = 100%	PA > 50%	Otherwise
C_3	NA = 100%	PA = 100%	Otherwise

Note: $NA = 1 - \frac{BMdocs}{CTdocs}$, PA = $\frac{BMdocs}{CTdocs}$

teraction with the 737 documents, we extracted a total of 121 candidate features and categorised them as shown in Table 2. Object features consisted of seven sub-categories, all of which extracted from click-through and bookmarked (BM) documents. Interaction features (Query Length, Rank of click-through URLs, Number of CT URLs so far, Time Spent so far and Number of queries submitted so far) were extracted from the transaction logs recorded by an experimental search interface.

To vary the level of agreement on document relevancy, we devised three conditions as shown in Table 3. We varied the level of agreement by increasing the amount of documents discarded from the classifiers. The first (C_1) was the most liberal condition where a document was judged (non-)relevant when more than half of click-through agreed. The documents which had a complete disagreement were removed in this condition. The second (C_2) was the same as C_1 excepts the criterion of non-relevant documents was strengthened to a complete agreement. Finally, C_3 used the documents whose relevancy was completely agreed on both relevant and non-relevant judgements. The varied levels of relevance judgements were used to train the classifiers and the effect of agreement was measured by the performance of relevancy prediction.

3. RESULTS AND IMPLICATIONS

The results of relevancy prediction are shown in Table 4. For the object features, the average performance of seven sub-categories is presented for simplicity. Sub-sampling was performed to keep the portion of relevant and non-relevant documents equal for the analysis, thus, the baseline performance was 50% in the table. As can be seen, the effect of highly agreed documents was little when all click-through documents were examined. This was consistent across the feature categories. However, a significant improvement was found in the prediction accuracy when multiple click-through documents were examined. In the object features, C_2 showed the best performance, suggesting that increasing the level of agreement for non-relevant documents can be effective. On the other hand, the interaction features

Table 4: Performance of relevancy prediction compared to a baseline performance (50%).

Feature category	CT Freq.	C_1	C_2	C_3
Object	All CT	+3.0	+2.7	+2.7
	CT>1	+6.6	+10.2	+7.8
	CT>2	+0.8	+12.0	+8.0
	Mean	+3.5	+8.3	+6.2
Interaction	All CT	+2.3	+2.6	+2.3
	CT>1	+2.2	+9.6	+10.9
	CT>2	+3.3	+14.1	+19.6
	Mean	+2.6	+8.6	+10.9
Overall mean		+3.0	+8.4	+8.6

further benefited from the increased level of agreement for both side of judgements. This suggests that an optimal level of agreement can differ across the category of features.

The results of the experiment showed that the classifiers improved the accuracy of relevancy prediction when the level of agreement was increased. This demonstrates that highly agreed documents can facilitate the mining of significant contextual features. An implication of this in the design of adaptive search models is that aggregated relevance information can be important for effective use of interaction data. For example, one can start to analyse the features of retrieved documents only when the frequency of click-through goes beyond a threshold. It is plausible that such a simple filtering can reduce the noise in the modelling of significant contexts. While this study was based on machine learning techniques, the finding might be applied to other approaches. Further investigation of our hypothesis is our future work.

4. REFERENCES

[1] E. Agichtein, et al. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th SIGIR Conference*, 3–10, 2006.

[2] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley Sons, New York, 1973.

[3] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM TOIS*, 23(2):147–168, 2005.

[4] L. J. H. Zhang and J. Su. Hidden naive bayes. In *Proceedings of AAAI-05*. AAAI Press, 919-924, 2005.

[5] P. Ingwersen and K. Järvelin. Information retrieval in context: IRiX. *SIGIR Forum*, 39(2):31–39, 2005.

[6] H. Joho and J. M. Jose. Slicing and dicing the information space using local contexts. In *Proceedings of the First IIiX Symposium*, 111–126, 2006.

[7] J. Pearl. *Probabilistic Reasoning with Intelligent Systems*. Morgan & Kaufman, San Mateo, 1988.

[8] E. Sormunen. Liberal relevance criteria of TREC -: counting on negligible documents? In *Proceedings of the 25th SIGIR Conference*, 324–330, 2002.

[9] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st SIGIR conference*, 315–323, 1998.

[10] G. I. Webb et al. Not so naive bayes: aggregating one-dependence estimators. *Mach. Learn.*, 58(1):5–24, 2005.