Multiple Approaches to Analysing Query Diversity

Paul Clough
University of Sheffield
Regent Court, 211
Portobello Street
Sheffield, S1 4DP
p.d.clough@
sheffield.ac.uk

Mark Sanderson
University of Sheffield
Regent Court, 211
Portobello Street
Sheffield, S1 4DP
m.sanderson @
sheffield.ac.uk

Murad Abouammoh
University of Sheffield
Regent Court, 211
Portobello Street
Sheffield, S1 4DP
m.abouammoh@
sheffield.ac.uk

Sergio Navarro
University of Alicante,
Departamento de
Lenguajes y Sistemas
Informaticos
snavarro@
dlsi.ua.es

Monica Paramita
University of Sheffield
Regent Court, 211
Portobello Street
Sheffield, S1 4DP
m.paramita@
sheffield.ac.uk

ABSTRACT
In this paper we examine user queries with respect to diversity: providing a mix of results across different interpretations. Using two query log analysis techniques (click entropy and reformulated queries), 14.9 million queries from the Microsoft Live Search log were analysed. We found that a broad range of query types may benefit from diversification. Additionally, although there is a correlation between word ambiguity and the need for diversity, the range of results users may wish to see for an ambiguous query stretches well beyond traditional notions of word sense.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Systems and Software --- performance evaluation.

General Terms
Measurement, Experimentation.

Keywords
Diversity, ambiguity.

1. INTRODUCTION
A system retrieving a spread of results for a user need or one that retrieves results across the interpretations of a query is said to be a system that promotes a *diverse ranking*. Offering diversity in search results can range from providing multiple distinct interpretations for ambiguous queries [1], to representing sub-topics within search results for queries with broad thematic scope [2]. In this paper we analyse query diversity based on testing two different (but complementary) approaches: click entropy [3] and query reformulations [4]. In this paper we examine queries from a real search log (Section 2) using multiple approaches to detecting diversity (Section 3). We report results for each approach and analyse query ambiguity (Section 4) before summarising (Section 5).

2. DATASET
The dataset used in this study was the Microsoft Live Search (MS) 2006 query log excerpt released as part of the Workshop on Web Search Click Data (http://research.microsoft.com/en-us/um/people/nickcr/wscd09/). It was composed of 14.9 million entries gathered over one month in 2006. The number of unique queries within the log was 6.62 million (from 5.43 million sessions). The logs contain a session ID, query text, date/time; if a URL in the search results was clicked, URL and rank were also recorded.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

3. DETECTING DIVERSITY
To investigate query diversity in the dataset, two techniques were used that potentially identify different forms of query diversity. *Click entropy* measures the spread of search results that a user (defined by anonymised code in the log) clicks on (higher scores reflect that users click on many results). Its use within our study was inspired by past work showing this measure to reflect novelty or diversity within ranked lists [3]. Only queries with 50 or more repetitions measured across at least 2 different sessions and with a total number of clicks greater than 5 were used. This left a set of 14,909 unique queries, which were found 4,181,582 times in the log. The presence of a query in a log followed quickly by a *reformulation* of the query suggests a failure in the search engine to cover user's interests in the initial search results. Researchers such as Radlinski and Dumais [4] showed that the more query-query reformulations executed, the higher the diversity. Therefore, this form of query pattern was extracted from the log. Reformulated queries were defined as queries found in the search logs where a user submitted at least two queries within a minute of each other that had at least one word in common.

4. RESULTS
Within the set of 14,909 *unique* queries, the level of click entropy was measured. It was found that queries with a 'high' click entropy value (>3) were almost entirely composed of information seeking queries; queries with a 'low' click entropy (≤3) were mainly navigational queries. Names of organizations, URLs (or URL fragments) were found to account for >95% of the low entropy queries (see Table 1). The queries with high diversity represented 18% of queries, and those with low diversity constituted 80.2% of the examined set; 1.8% of the set had an entropy of zero.

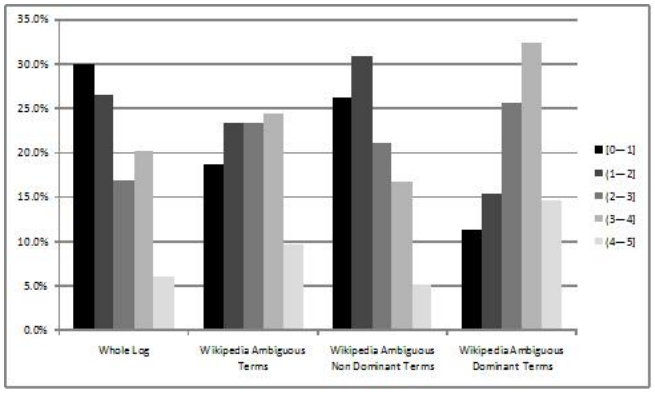
Entropy	URL	Organization name	Other
≤3	44.4%	48.5%	7.1%
=0	75.6%	22.2%	2.2%

Table 1: Distribution of different forms of navigational queries for different click entropy values.

To explore *ambiguity* within the 14,909 commonly repeated queries of the MS search log, a list of all ambiguous terms (words and phrases) in WordNet and Wikipedia was generated, along with a list of the number of senses for each term. For Wikipedia, the disambiguation pages were used as a source for ambiguous terms (see [5]).

For all terms occurring as a *complete* query, a click entropy value was calculated. The number of ambiguous words found in the 14,909 queries from WordNet was 1,592 and in Wikipedia 2,452 (with an intersection of 1,189 queries). Using Pearson's correlation coefficient revealed little significant correlation between the number of senses and the click entropy (r=-0.02 for WordNet; r=0.07 for Wikipedia). On the basis of these results, it would appear that the sense counts in these sources are of little value in predicting the level of click entropy and consequently diversity. Furthermore, popularly-cited ambiguous queries such Jaguar [6], Java [7] and bank [8] had low click entropies (2.73, 3.73 and 1.68 respectively) indicating users were not typically clicking on a wide range of results for these queries.

Fig 1: Click entropy values for different subsets of ambiguous queries (values in charts are in the same order as key)



It is possible to extract two distinct forms of ambiguous term from Wikipedia: those with a dominant sense and those without. It was decided to split the 2,452 ambiguous Wikipedia words/phrases into two sets: dominant (1,202) and non-dominant (1,250). Fig 1 shows click entropy for different subsets of ambiguous queries and we observe that the distributions vary depending on the two kinds of ambiguous term: queries with dominant senses have a broad spread of entropy values (perhaps indicative that dominant sense topics in Wikipedia usually cover a broad topic); those with non-dominant senses are similar to values for the entire log.

To examine this further, the Wikipedia article describing each of the 1,202 terms with a dominant sense was downloaded and size of article (in words) computed. Using Pearson's correlation coefficient we found a significant correlation between the size of the article and click entropy for the term (r=0.23; p<0.05). Whilst not strong, there appears a positive correlation between the size of the article and users' preference for selecting a diverse set of results. Assuming that the size of a Wikipedia article is indicative of the breadth of the topic, the broader the topic, the greater the diversity observed in user click behaviour.

For the reformulation queries, it was found that 17% of all queries were part of some kind of reformulation with 7% being the initial queries (*pre-formulated queries*) and 10% the reformulations. (There were more reformulations as queries could be successively reformulated.) The distribution of reformulations was computed, showing that 76.25% of the queries were reformulated only once, 15.98% twice and 4.70% three times.

Finally, we combined estimates from the click entropy and reformulation analysis (Table 2), thresholding the number of times queries were repeated in the logs. Results showed between 9.5% and 16.2% of queries submitted to MS Live Search could benefit from having diversity applied. It is also noticeable that a relatively small overlap between the queries identified through the two approaches exists. This confirms the need for further exploration into multiple approaches to analyse query diversity.

Num Repetitions≥	50	50	10
Num Reformulations≥	50	10	10
Only Click entropy	4.55%	2.86%	4.63%
Only Reformulation	4.33%	9.17%	9.16%
Entropy & Reformulation	0.60%	2.39%	2.40%
Total Diverse	9.49%	14.43%	16.20%
Non Diverse	90.51%	85.57%	83.80%

Table 2: Lower bound estimates of number of queries that might benefit from diverse search output.

5. CONCLUSIONS AND FUTURE WORK

This paper discusses an analysis of query diversity based on two different approaches: click entropy and reformulated queries. Each approach was used with the intent of investigating diversity in the 14.9 million query Microsoft Live Search log. The analysis showed that at least 9.5%-16.2% of queries could benefit from diversification. Although there has been research focusing on promoting diversity for person names (e.g. WEPS) and spatial proximity, there has been less focus on other types of query. Ambiguity, as defined in thesauri or large community-built resources like Wikipedia, does not appear to have a strong correlation with the forms of diverse queries observed in the query logs (although there is some indication that the length of a Wikipedia article is correlated with query result diversity). More work needs to be carried out to establish the implications of diverse results with interfaces, end users and their information seeking needs and behaviours.

6. ACKNOWLEDGMENTS

Work was partly supported by the following EU-funded projects: TrebleCLEF (Grant agreement #215231), TRIPOD (Contract #045335) and QALL-ME (FP6-IST-033860). Also supported by the Spanish Government funded project TEXT-MESS (TIN-2006-15265-C06-01).

7. REFERENCES

[1] R. Agrawal et al., "Diversifying search results," *In Proceedings of WSDM'09,* 2009, pp.5-14.

[2] E. Vee et al., "Efficient Computation of Diverse Query Results," *In Proceedings of ICDE'08*, 2008, pp.228-236.

[3] R. Song et al., "Identification of ambiguous queries in web search," *Information Processing and Management*, 2008.

[4] F. Radlinski and S. Dumais, "Improving personalized web search using result diversification," *In Proceedings of SIGIR'06,* 2006, pp. 691-692.

[5] M. Sanderson, "Ambiguous queries: test collections need more sense," *In Proceedings SIGIR'08*, 2008, pp. 499-506.

[6] C.L.A. Clarke et al., "Novelty and diversity in information retrieval evaluation," *In Proceedings of SIGIR'08*, 2008, pp. 659-666.

[7] J. Bai and J.Y. Nie, "Adapting information retrieval to query contexts," *Information Processing and Management*, vol. 44, 2008, pp. 1901-1922.

[8] R. Richardson and A.F. Smeaton, "Using WordNet in a Knowledge-Based Approach to Information Retrieval," *In Proceedings of the BCS-IRSG Colloquium*, Crewe, 1995.