

Estimating the Value of Automatic Disambiguation

Paul Thomas
Department of Computer Science
Australian National University
Canberra, Australia
paul.thomas@anu.edu.au

Tom Rowlands
CSIRO ICT Centre
Canberra, Australia
tom.rowlands@csiro.au

ABSTRACT

A common motivation for personalised search systems is the ability to disambiguate queries based on some knowledge of a user's interests. An analysis of log files from three search providers, covering a range of scenarios, suggests that this sort of disambiguation would be of marginal use for more specialised providers but may be of use for whole-of-Web search.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*search process*

General Terms

Measurement

Keywords

Disambiguation, personalisation

1. INTRODUCTION

Personalisation techniques in information retrieval are often motivated by disambiguation: if a query term is ambiguous, a personalised system can use knowledge of the user to determine the term's correct meaning and thus improve precision (e.g. [2, 6, 8, 9, 10]). There are also implementations of personalised search in commercial whole-of-Web search engines, and these implementations appear to carry out some disambiguation. The question we address here is: is disambiguation likely to provide a real benefit in search, and if so in which scenarios?

We have examined log files from three search providers (the public website of a company; a portal for government websites; and a large whole-of-Web search engine) to count the number of sessions in which users issued a query to disambiguate an earlier one, for example following “jaguar” with “jaguar cars”. This provides an estimate of the number of queries which are:

1. ambiguous in one or more terms,
2. not already answered well by the search provider, and
3. important enough for the user to warrant the effort of reformulating the query.

Our count therefore provides an estimate of the proportion of sessions which could be improved by automatic disambiguation, and an indication of the scenarios where the technique may be most useful.

2. ANALYSIS

Logs were acquired from three search providers, covering a range of scales: one was a medium-sized company, one a portal to a range of government websites, and one a whole-of-Web search engine. Each log included the text, time, and a user identifier or client IP address for each query.

For each provider, the logged data was divided into sessions: a “session” was defined as all queries from the same user or client IP address with not less than five minutes inactivity. Any session with only one query could not include explicit disambiguation and was removed. A uniform subsample of at least 750 of the remaining sessions was taken and given to two judges for manual classification. Each judge marked those sessions which appeared to contain an instance of explicit disambiguation.

The company's search engine indexes around 22,000 pages. Log files used covered 5 days and 11,733 sessions, of which 939 (8%) had more than one set of terms and were candidates for judging. The government portal is significantly larger, with 2,300,000 pages and 59,717 sessions across 24 days; 1720 sessions (3%) were candidates for judging. Logs from the whole-of-Web engine covered three months and 338,775 sessions of which 96,788 (29%) were candidates.

Explicit disambiguation generally took the form of extra terms being added to narrow the scope of queries, but on some occasions terms were replaced with near synonyms. Examples from the logs included:

holidays	→	public holidays
prefect	→	prefect of melbourne
methane emissions	→	methane emissions restrictions
report on wages	→	pharmacist salary data

Results are summarised below. “Explicit disamb.” is the proportion of all sessions which appear to include an instance of explicit disambiguation; these are the sessions for which automatic disambiguation would be a clear advantage. “Inter-judge agreement” is the Jaccard similarity between the judges' assessments, $|J_1 \cap J_2|/|J_1 \cup J_2|$, where J_1 and J_2 are the sets of sessions marked by each judge. We note that agreement of 71–88% is in line with results of many TREC experiments [11].

Search provider	Explicit disamb.	Inter-judge agreement
Single company	1.3%	85%
Govt. portal	1.7%	88%
Whole-of-Web	6.5%	71%

The similarity between the results for the single company and government portal is unexpected: it would seem reasonable to assume that with the greater range of subject matter in the government portal there would be more scope for ambiguity (criterion 1 above) and that government business would be more important to users (criterion 3). Since the proportion of sessions with more than one query is so low for this provider, we believe the difference may be in the quality of the search engine and in the nature of common queries.

The final figures for the single company and government portal are low (1.3% and 1.7% respectively). While these results do indicate a potential benefit from automatic disambiguation, they also suggest the impact of such a system will be small, and that the overall improvement may not be worth the additional effort and additional risk of error.

The whole-of-Web result of 6.5%, however, suggests that automatic disambiguation would be helpful in general Web search. This is reinforced by the observation that a full 29% of sessions in our data involve more than one query. The difference observed between whole-of-Web and more focussed search could be explained by several factors:

- Coverage of the Web is less clear than that of a single website, so it may seem more likely that a reformulated query will work.
- The range of topics covered by the Web is huge, so any one query is more likely to be ambiguous.
- Web search users may be more vague about their information need — an informed user may be more likely to instead use a specialised portal. Portal users may also be more experienced and be aware of likely ambiguity.
- The number of documents available in such a large corpus may mean one dominant interpretation is more likely to flood others out, leaving users less likely to be satisfied if their query is even slightly ambiguous.

Finally, we note that the figures above represent the gain available only if we can successfully disambiguate in all cases. With realistic success rates of about 60% [7], the benefit of automatic systems would be further limited.

3. RELATED WORK

In contrast to the investigation reported here, research on disambiguation has typically used standard test corpora and artificial queries. Following an investigation using artificial terms in the Reuters corpus [5], Sanderson concluded that automatic disambiguation is likely to be of only small benefit to information retrieval systems and then only if the disambiguation is of high quality. Stokoe, Oakes, and Tait broadly agree [7], finding an improvement on TREC Web Track tasks but only of very small degree (around 0.03 improvement in R-precision). Similar small benefits have been noted by other studies over TREC-like data (e.g. [1, 3, 4]),

but larger benefits are suggested with Web search [9]. This is consistent with our findings.

As well as systems which use disambiguation to improve performance over all users, a number of systems have been built to personalise a system by performing disambiguation (eg [2, 6, 10]). To the best of our knowledge, the improvements due to disambiguation have not been systematically measured for these “real world” systems.

4. CONCLUSIONS

Personalisation systems motivated by disambiguation assume there are enough instances of ambiguous and poor queries to be worthwhile. We find explicit disambiguation in only 1.3–1.7% of sessions with commercial and government search providers, but 6.5% with whole-of-Web search. This suggests that automatic disambiguation will only realise a small benefit in more focussed environments but may be somewhat helpful in Web search.

5. ACKNOWLEDGEMENTS

We thank David Hawking for useful feedback on these experiments. We would also like to thank the (anonymous) providers of log files for their support.

6. REFERENCES

- [1] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarrán. Indexing with WordNet synsets can improve text retrieval. In *Proc. COLING/ACL Workshop on Usage of WordNet for Natural Language Processing*, 1998.
- [2] G. Koutrika and Y. Ioannidis. A unified user profile framework for query disambiguation and personalisation. In *Proc. Workshop on New Tech. for Personalized Info. Access*, 2005.
- [3] S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *Proc. ACM SIGIR*, 2004.
- [4] R. Mihalcea and D. Moldovan. Semantic indexing using WordNet senses. In *Proc. ACL workshop on Recent Advances in NLP and IR*, 2000.
- [5] M. Sanderson. Word sense disambiguation and information retrieval. In *Proc. ACM SIGIR*, 1994.
- [6] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proc. CIKM*, 2005.
- [7] C. Stokoe, M. P. Oakes, and J. Tait. Word sense disambiguation in information retrieval revisited. In *Proc. ACM SIGIR*, 2003.
- [8] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proc. WWW*, 2004.
- [9] J. Teevan, S. T. Dumais, and E. Horvitz. Beyond the commons: Investigating the value of personalizing web search. In *Proc. Workshop on New Tech. for Personalized Info. Access*, 2005.
- [10] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proc. ACM SIGIR*, 2005.
- [11] E. Voorhees and D. Harman. Overview of the fifth Text REtrieval conference. In *Proc. TREC*, 1997.