A User Study on Snippet Generation: Text Reuse vs. Paraphrases

Wei-Fan Chen[1] Matthias Hagen[2] Benno Stein[1] Martin Potthast[3]

[1]Bauhaus-Universität Weimar
<first>.<last>@uni-weimar.de

[2]Martin Luther University Halle-Wittenberg
matthias.hagen@informatik.uni-halle.de

[3]Leipzig University
martin.potthast@uni-leipzig.de

ABSTRACT

The snippets in the result list of a web search engine are built with sentences from the retrieved web pages that match the query. Reusing a web page's text for snippets has been considered fair use under the copyright laws of most jurisdictions. As of recent, notable exceptions from this arrangement include Germany and Spain, where news publishers are entitled to raise claims under a so-called ancillary copyright. A similar legislation is currently discussed at the European Commission. If this development gains momentum, the reuse of text for snippets will soon incur costs, which in turn will give rise to new solutions for *generating truly original snippets*. A key question in this regard is whether the users will accept any new approach for snippet generation, or whether they will prefer the current model of "reuse snippets." The paper in hand gives a first answer. A crowdsourcing experiment along with a statistical analysis reveals that our test users exert *no significant preference* for either kind of snippet. Notwithstanding the technological difficulty, this result opens the door to a new snippet synthesis paradigm.

ACM Reference Format:
Wei-Fan Chen, Matthias Hagen, Benno Stein, and Martin Potthast. 2018. A User Study on Snippet Generation: Text Reuse vs. Paraphrases. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210149

1 INTRODUCTION

Snippets are an essential part of a search results page: they incite users to view (to click) or to skip viewing a retrieved document. Already in 1991, Pedersen, Cutting, and Tukey [15] proposed query-biased snippets, and they have proven useful until today [23, 25, 26]. The more surprising appears Google's recent decision to remove snippets altogether from its redesigned news portal, without offering any explanation.[1] Recall that Google has notoriously been "questioning the unquestioned," subjecting virtually every detail of its search interfaces to A/B tests. If this policy has not been changed, can the redesigned news portal be interpreted as evidence that snippets are not so useful after all (in the domain of news search)? A more plausible explanation can be found in the changing interpretation of the copyright: publishers from all over the world are now raising claims for compensation for displaying text extracted from their news articles. Especially in Europe their lobbying for political support was successful: an ancillary copyright for news publishers, which basically exempts their intellectual property from fair use, has been passed into law in Germany and Spain, and it is currently discussed as part of an EU-wide copyright reform. In light of these developments, the removal of snippets from Google News appears as an act of anticipatory obedience.

Inasmuch as today's information economy on the web is financed by advertisements, the welfare of publishers partaking in this ecosystem depends on consumers visiting their web pages. This is also true for a large portion of the revenue of news publishers; some well-known publishers even stopped printing newspapers. It is hence no surprise that news publishers protect their online assets more fiercely than they used to do. News publishers form a well-organized business community with traditionally strong ties to politics and public opinion, yet, other online communities may follow. Even Wikipedia's contents are regularly lifted onto the search results pages of many search engines, depriving the encyclopedia of visitors, which may have contributed to the ongoing decline of active Wikipedia editors since 2007 [13, 21]. Commercial search engines, whose operations remained unchallenged in this respect for more than two decades, may therefore face a turn.

Are information scientists forced to pick a side? Probably not, since it is not our business to protect business models, neither that of search engines nor that of publishers. Rather, we should uphold the vision of developing the "perfect" information system, which is what the information society needs. When it comes to snippets, text reuse has been popular because it is easy. Looking forward, Potthast et al. [17] propose deep-learning-based text generation as a promising, yet more difficult alternative. Moreover, users' expectations may include reuse snippets. The paper in hand debunks this notion: we investigated the users' preferences, showing that, with some reservations, the majority of users does not prefer traditional reuse snippets over paraphrased versions, or vice versa.

2 RELATED WORK

Snippet generation is a variant of extractive summarization, where the summaries are biased toward queries. Luhn, the inventor of term frequency weighting, was one of the earliest contributors [2, 11]. Tombros and Sanderson [23] ascertained the importance that snippets relate to a user's query, while Brin and Page [3] implemented query-biased snippets for the first version of Google. White et al. [25, 26] found that snippets should be re-generated based on implicit relevance feedback when a user returns to a search results page. To speed up snippet generation, Turpin et al. [24] evaluate software architectures based on compressed data structures and RAM caching. Bando et al. [1] ask humans to manually create reuse snippets, comparing the results to machine-generated reuse snippets. They observe that in about 73% of cases humans select the same pieces of text as machines. Savenkov et al. [19] survey

[1]www.blog.google/topics/journalism-news/redesigning-google-news-everyone

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210149

approaches regarding the evaluation of snippet generation, suggesting automated evaluation approaches and A/B testing. Thomaidou et al. [22] consider the special case of snippet generation for ads that are shown on search results pages. Further research has been invested into studying how the length of snippets affects the perceived search result quality on desktops [9, 12] as well as on mobile devices where screen space is limited [10]. Eye-tracking studies have been conducted to determine to what parts of a search results page users pay most attention [5, 7]; unsurprisingly, snippets play a major role. Finally, reuse snippets are also generated in XML retrieval [8] and semantic web search [16].

The companion task to extractive summarization is *abstractive* summarization [6], where summaries are not restricted to text reuse. Though abstractive summarization is a long-standing task in the natural language generation community, it has not been considered for snippet generation yet. In their user study, Bando et al. [1] come close, using manually written non-reuse snippets as a gold standard to evaluate automatically generated reuse snippets. It was shown that humans pay attention to the same parts of a document when (1) manually composing a snippet compared to when (2) selecting sentences for a snippet. Recently, neural network models have made great progress toward the task of abstractive summary generation [4, 14, 18, 20], which renders snippet synthesis feasible if the lack of large-scale training data can be overcome.

3 USER STUDY DESIGN

Within three crowdsourcing tasks, we first acquired paraphrases of reuse snippets, and then experimentally determined which kind workers prefer when given a pair, and which kind is more useful to spot relevant results on a search results page.

3.1 Crowdsourcing Paraphrase Snippets

Given Bando et al.'s [1] insight of the high overlap of human reuse / paraphrase snippets to machine-generated reuse snippets, paraphrase snippets represent a sufficient substitute for true original snippets for our user study. To maximize diversity of the set of pairs of reuse snippets and corresponding paraphrase snippets, we resort to crowdsourcing. Using the 150 topics provided for the TREC Web tracks 2009–2011, each topic's query has been submitted to Google's custom search API[2] to obtain high-quality search results. The top-5 search results of each query were collected, including title, URL, and Google's reuse snippet for a total of 750 snippets. Since Google's snippet generator sometimes shortens sentences to enforce a maximum snippet length (indicated by ellipses), we recovered the complete sentences from the linked pages. For crowdsourcing we relied on Amazon's Mechanical Turk (AMT), where we offered the task to manually paraphrase the reuse snippets collected. The worker instructions were to significantly rewrite a given snippet while maintaining its length and without removing important named entities, phrases, or quotes (e.g., "to be or not to be"). To foreclose easy cheating, copy and paste was disabled in the AMT interface. Each of the 750 reuse snippets was assigned to two different workers for paraphrasing (i.e., we repeated our experiment twice in a row to test its reliability). Submitted paraphrases were reviewed, rejecting those with lacking changes or poor grammar, resulting in 1,500 pairs of reuse and paraphrase snippets.

[2]https://developers.google.com/custom-search/

3.2 Snippet Preference

To assess snippet preference, we recruit 5 workers for each pair of reuse / paraphrase snippets to judge which of the two they would prefer for the given query and web page. The task interface showed instructions, a search box with the topic's query, the pairs of snippets side by side, formatted like standard search results, the associated web page in a frame below, and a text field to enter an assessment justification. Workers were asked to assign one of four labels: snippet 1 or 2 is better, both are good, or both are bad. To avoid order bias, the positions of reuse and paraphrase snippets were randomized. After collecting all judgments, we tallied the scores as follows: a reuse or paraphrase snippet gets 1 point if a worker judged it to be better, both get a point if a worker judged that both are good, neither get a point otherwise.

The worker pool of AMT has been known to comprise dishonest workers, threatening the reliability of our study. We took several precautions: each worker judged at most two pairs of snippets ensuring diversity, and submissions were rejected if workers spent insufficient time, too much time, or if they failed to provide sensible explanations for their judgments, resulting in 4,235 individual workers and 7,500 accepted annotations. Only workers having at least 80% acceptance rate and at least 100 successful assignments were invited. Furthermore, we conducted control experiments with respect to the variables snippet source, preference bias, snippet length, and random pairings to check how workers are affected.

3.3 Snippet Usefulness

To obtain implicit feedback on a snippet's usefulness for spotting relevant search results, another group of workers judged the relevance of a search result to a query given different page configurations. The queries, corresponding web pages, and relevance scores were obtained from the topics used at the TREC Web tracks 2013–2014, which were based on the ClueWeb12. For each topic, we tried to collect 3 web pages judged as relevant and 3 judged as irrelevant that are still available on the live web and whose contents correspond to that found in the ClueWeb12.[3] For 29 topics, we were able to collect the desired set of 6 web pages. Following the aforementioned procedures, we collected reuse snippets using Google's custom search API, and paraphrases of them via crowdsourcing.

Workers were then exposed to search results pages comprising 3 results (1) with reuse snippets, (2) with paraphrase snippets, (3) without snippets (only titles and URLs), (4) with reuse snippets only (no titles or URLs), or (5) with captcha-style snippets to ensure workers read the snippets. In the latter case, the snippets just stated whether a result was supposed to be relevant or irrelevant. A search results page could contain 0 up to 3 relevant web pages. For mixtures of relevant and irrelevant pages, we tested all permutations of search result orderings. For each ordering, three workers provided labels, yielding a total of 10,440 annotations (29 topics, 8 relevance settings (0–3 results relevant), 5 snippet conditions (reuse, paraphrase, etc.), 3 results per search results page, and 3 annotators each). Each worker judged search results pages of 5 different topics based on the given information. To ensure annotation quality, we rejected results from workers who did not pass the captcha snippets, resulting in 546 individual workers in this experiment.

[3]Topics from the previous TREC Web tracks were omitted, since they are based on the ClueWeb09 which is insufficiently represented on today's web.

Table 1: Left: Distribution of judgments; 1,500 pairs of (reuse, paraphrase) snippets at 5 assessors each yield 7,500 judgments. Middle: Average scores (number of votes) of reuse and paraphrase snippets with p values for a paired t-test, bold font indicating significance ($p < 0.05$); the two row groups correspond to two repetitions of the experiment. Right: Average scores of pairs of snippets grouped by different aspects, with associated p values and one row group per experiment repetition.

Assessment	**Judgments**	
	absolute	relative
Reuse better	2,731	36.41%
Paraphrase better	2,652	35.36%
Both good	1,537	20.49%
Both bad	580	7.74%
Total	7,500	100.00%

Experiment	**Reuse**	**Paraphrase**	***p*-value**
all	3.06	2.97	0.51
Wikipedia	3.31	2.58	**0.00**
Non-Wikipedia	2.75	2.85	0.31
all	3.05	2.94	0.43
Wikipedia	3.18	2.64	**0.01**
Non-Wikipedia	2.77	2.82	0.58

Experiment (par. = paraphrase, re. = reuse)	**(•,)**	**(,•)**	***p*-value**
(better, worse)	3.91	1.71	**0.00**
((better-par., worse-re.), (better-re., worse-par.))	2.85	2.78	0.41
(long, short)	2.91	2.70	**0.01**
((long-par., short-re.), (long-re., short-par.))	2.82	2.81	0.90
(better, worse)	3.89	1.75	**0.00**
((better-par., worse-re.), (better-re., worse-par.))	2.87	2.78	0.23
(long, short)	2.99	2.61	**0.00**
((long-par., short-re.), (long-re., short-par.))	2.79	2.83	0.60

4 USER STUDY ANALYSIS

We conducted a careful statistical analysis of the crowdsourced snippet judgments. The snippet preference experiment rests on the hypothesis that users do not *consciously* care whether or not snippets reuse content from linked web pages as long as they are semantically equivalent. If true, there should be no statistically significant difference in terms of user preference. The snippet usefulness experiment rests on the hypothesis that users are not *unconsciously* negatively affected by paraphrase snippets. If true, users identify relevant web pages either way and there should be no statistically significant differences between the two kinds of snippets.

4.1 Descriptive Statistics

The reuse snippets collected comprise an average of 1.9 sentences and 41.1 words; the longest snippet has 6 sentences and 122 words, the shortest one 1 sentence and 14 words. The paraphrase snippets comprise an average 2.2 sentences and 40.5 words; the longest snippet has 6 sentences and 132 words, the shortest one 1 sentence and 9 words. The paraphrase snippets are significantly longer at the sentence level ($p < 0.05$), but neither are significantly shorter nor longer at the word level ($p > 0.05$). A reason might be that workers tended to split long sentences while paraphrasing. The workers spent an average 220.5 seconds to paraphrase a snippet at a maximum of 895 seconds (38 words) and a minimum of 14 seconds (also 38 words), an average of 49 seconds to judge a pair of snippets, and of 17 seconds to judge a web page's relevance when viewing a search results page. The inter-annotator agreement Fleiss κ for the snippet preference experiment, presuming the four labels to be independent, was 0.37, indicating a fair agreement, and 0.77 for the snippet usefulness experiment, indicating a substantial agreement.

4.2 Snippet Preference

Judgment distribution. Table 1 (left) shows the distribution of judgments. Recall that workers were unaware which snippet is which; their judgments were mapped to the ground truth afterwards. The amounts of judgments for *Reuse better* and *Paraphrase better* are roughly equal and about a quarter of workers had no preference (*Both good* plus *Both bad*). Only 580 pairs of snippets (7.74%) were judged *Both bad*, showing a high overall snippet quality.

Reuse snippets vs. paraphrase snippets. To check for snippet preferences, we performed a paired t-test on the pairs of reuse and paraphrase snippets. Since we repeated our experiment, collecting two different paraphrase snippets for each of the 750 reuse snippets, we can attest that our results can be replicated under the same conditions: the rows *all* in Table 1 (middle) show the results for the two repetitions. While the absolute average scores (number of votes) achieved by paraphrase snippets are slightly smaller than those of reuse snippets, no statistically significant difference was measured, given pretty high p values of 0.51 and 0.43, respectively.

Wikipedia snippets vs. non-Wikipedia snippets. From the 750 search results, 260 refer to Wikipedia articles. Considering only this subset, the rows *Wikipedia* in Table 1 (middle) show that users significantly prefer reuse snippets over paraphrases, which is not the case for non-Wikipedia results. The effect sizes under Cohen's d are small to medium (0.51 and 0.31, respectively). Upon review of Wikipedia snippet pairs, many of the reuse snippets have an a-priori high writing quality, and it may have been difficult for the average AMT worker to compete with that.

Preferred snippets vs. unpreferred snippets. To quantify the difference between snippets preferred by users (better) to those not preferred (worse), we reordered the snippet pairs accordingly, disregarding ties, and then applied a paired t-test. The rows *(better, worse)* in Table 1 (right) show the results for each repetition of our experiment. As can be seen, there is an average 2.2 score difference between them, rendering the differences significant. However, when comparing the groups of snippet pairs (better-reuse, worse-paraphrase) with (better-paraphrase, worse-reuse), p-values of 0.41 and 0.23 indicate that snippet preference is independent of whether they are reused or paraphrased.

Long snippets vs. short snippets. We further investigated if snippet length affects preference (rows *(long, short)* of Table 1 (right). A snippet belongs to the "long" snippets if it is the longer one of a pair, and to "short" snippets otherwise. On average, the long snippets have 44.7 words and the short snippets have 36.7 words. Our findings corroborate those of Maxwell et al. [12], namely that users prefer longer snippets. In fact, many of the assessment justifications from our workers support this finding. Again, when comparing the groups of snippet pairs (long-paraphrase, short-reuse) with (long-reuse, short-paraphrase), p-values of 0.90 and 0.60 indicate that the dimensions length and reuse are independent.

Reuse snippets vs. unrelated snippets. As a control experiment to ascertain worker diligence, pairs of reuse snippets and unrelated snippets were shown to workers, where a given reuse snippet was paired with a random reuse snippet of a different web page of a different query. Of 1,500 judgments collected, workers preferred the snippet matching the query in 85% of the cases, confirming this experiment's setup validity.

Table 2: F-scores of the snippet usefulness experiment indicating whether annotators correctly spot relevant search results under different search results page snippet conditions.

	Reuse	Paraphrase	No snippet	Snippet only	Random
F-score	67.64	64.61	63.65	60.16	50.00

4.3 Snippet Usefulness

This experiment questioned whether users can identify relevant pages given different kinds of snippets—one of the key tasks snippets should support. A search result is labeled as relevant if more than half of the workers label it as relevant, and irrelevant otherwise. In Table 2 we show the F-score of the crowdsourced judgments based on snippets compared with the ground truth relevance labels obtained from the TREC assessors. The numbers of overall shown relevant and irrelevant web pages were balanced, such that random guessing yields a baseline F-score of 50%. In the captcha-style setting where the snippets just explicitly state that a web page is relevant / irrelevant, the workers achieved an F-score of 100% (since we excluded those who did not succeed in these check instances). As for the other snippet conditions, we find that although reuse snippets achieve the highest F-score (helping users best to judge a result's relevance), the performance of paraphrase snippets is not significantly worse ($p = 0.28$). Showing only reuse snippets (without titles or URLs) achieves the lowest F-score; no snippets (only title and URL) is better than showing only snippets, confirming that title and URL do play an important role. All settings are significantly better than random guessing. Otherwise, only reuse snippets are significantly better than showing only snippets ($p < 0.05$). The remaining pairings are not significantly different, corroborating that paraphrase snippets are no worse than reuse snippets.

We conclude that the combination of snippet, title, and URL is crucial to identify relevant web pages on search results pages, regardless of whether snippets are reused or paraphrased. Nevertheless, there is room for improvement given the results obtained from showing the "perfect" snippet, which reveals the relevance of a web page (our captcha setup). The finding that both reuse and paraphrase snippets are useful supports our claim that paraphrase snippets can replace reuse snippets in future information systems.

4.4 Reproducibility

User studies need to be reproduced in order to determine whether the results of previous studies on a given problem of interest generalize and that they were not due to unidentified confounding variables or accidental flaws in the study setup. This pertains particularly to first-time studies, since only an independent reproduction will provide for sufficient confidence that the results obtained are valid and that they may generalize. Since our study is the first of its kind that provides an answer to the question whether reusing text for snippet generation is a necessity, or whether also paraphrased snippets are sufficient, we expect that sooner or later it will have to be reproduced for its results to be corroborated. The reproducibility of a user study rests with a clear description of its setup, which we tried our best to provide. But maybe even more so it rests with access to data, code, and supplementary material that was gathered throughout the study. To ensure the reproducibility of our results, we provide all data collected and the associated code open source.[4]

[4]https://github.com/webis-de/SIGIR-18

5 CONCLUSION AND FUTURE WORK

Our user study shows that reuse snippets have no significant advantage over paraphrase snippets, so that, if done right, snippet synthesis without relying on text reuse will not result in a worse user experience. However, generating coherent snippets, even if "just" paraphrasing reuse snippets, is still beyond today's text generation capabilities. Even deep learning cannot yet be applied out of the box, since the substantial amounts of training data required are still missing. In future work, we will focus on compiling a suitable training dataset. Regarding the political debate on whether reuse snippets should be regulated, it can already be said that if the pressure on commercial search engines is further increased, the demand for snippet synthesis technology may soon outweigh the costs of developing it. By the time new copyright laws are enacted, new technologies may have already superseded them.

REFERENCES

[1] L. L. Bando, F. Scholer, and A. Turpin. 2010. Constructing Query-biased Summaries: A Comparison of Human and System Generated Snippets. In *Proc. of IIiX.* 195–204.
[2] P. B. Baxendale. 1958. Machine-Made Index for Technical Literature - An Experiment. *IBM Journal of Research and Development* 2, 4 (1958), 354–361.
[3] S. Brin and L. Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30, 1-7 (1998), 107–117.
[4] S. Chopra, M. Auli, and A. M. Rush. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proc. of NAACL/HLT.*
[5] E. Cutrell and Z. Guan. 2007. What are you Looking for?: An Eye-tracking Study of Information Usage in Web Search. In *Proc. of CHI.* 407–416.
[6] M. Gambhir and V. Gupta. 2017. Recent Automatic Text Summarization Techniques: A Survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.
[7] L. A. Granka, T. Joachims, and G. Gay. 2004. Eye-tracking Analysis of User Behavior in WWW Search. In *Proc. of SIGIR.* 478–479.
[8] Y. Huang, Z. Liu, and Y. Chen. 2008. Query biased Snippet Generation in XML Search. In *Proc. of SIGMOD.* 315–326.
[9] M. Kaisser, M. A. Hearst, and J. B. Lowe. 2008. Improving Search Results Quality by Customizing Summary Lengths. In *Proc. of ACL.* 701–709.
[10] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H. Yoon. 2017. What Snippet Size is Needed in Mobile Web Search?. In *Proc. of CHIIR 2017.* 97–106.
[11] H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2, 2 (1958), 159–165.
[12] D. Maxwell, L. Azzopardi, and Y. Moshfeghi. 2017. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *Proc. of SIGIR.* 135–144.
[13] C. McMahon, I. Johnson, and B. Hecht. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *Proc. of ICWSM.*
[14] R. Nallapati, B. Zhou, C. Nogueira dos Santos, Ç. Gülçehre, and B. Xiang. 2016. Abstractive Text Summarization using Sequence-to-Sequence RNNs and Beyond. In *Proc. of CoNLL.*
[15] J. Pedersen, D. Cutting, and J. Tukey. 1991. Snippet Search: A single phrase approach to text access. In *Proc. of the 1991 Joint Statistical Meetings.*
[16] T. Penin, H. Wang, T. Tran, and Y. Yu. 2008. Snippet Generation for Semantic Web Search Engines. In *Proc. of ASWC.* 493–507.
[17] M. Potthast, W. Chen, M. Hagen, and B. Stein. 2018. A Plan for Ancillary Copyright: Original Snippets. In *Proc. of NewsIR*, Vol. 2079. CEUR-WS.org, 3–5.
[18] A. M. Rush, S. Chopra, and J. Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proc. of EMNLP.*
[19] D. Savenkov, P. Braslavski, and M. Lebedev. 2011. Search Snippet Evaluation at Yandex: Lessons Learned and Future Directions. In *Proc. of CLEF 2011.* 14–25.
[20] A. See, P. J. Liu, and C. D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proc. of ACL.*
[21] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. 2009. The singularity is not near: slowing growth of Wikipedia. In *Proc. of WikiSym.*
[22] S. Thomaidou, I. Lourentzou, P. Katsivelis-Perakis, and M. Vazirgiannis. 2013. Automated Snippet Generation for Online Advertising. In *Proc. of CIKM.* 1841–1844.
[23] A. Tombros and M. Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proc. of SIGIR.* 2–10.
[24] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams. 2007. Fast Generation of Result Snippets in Web Search. In *Proc. of SIGIR.* 127–134.
[25] R. White, I. Ruthven, and J. M. Jose. 2002. Finding Relevant Documents Using Top Ranking Sentences: An Evaluation of Two Alternative Schemes. In *Proc. of SIGIR.* 57–64.
[26] R. White, I. Ruthven, and J. M. Jose. 2002. The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. In *Proc. of ECIR.* 93–109.