Evaluating Whole-Page Relevance

Peter Bailey, Nick Craswell, Ryen W. White,
Liwei Chen, Ashwin Satyanarayana, and S.M.M. Tahaghoghi
{pbailey, nickcr, ryenw, liweich, assatya, stahagh}@microsoft.com
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052 USA

ABSTRACT

Whole page relevance defines how well the surface-level representation of all elements on a search result page and the corresponding holistic attributes of the presentation respond to users' information needs. We introduce a method for evaluating the whole-page relevance of Web search engine results pages. Our key contribution is that the method allows us to investigate aspects of component relevance that are difficult or impossible to judge in isolation. Such aspects include component-level information redundancy and cross-component coherence. The method we describe complements traditional document relevance measurement, affords comparative relevance assessment across multiple search engines, and facilitates the study of important factors such as brand presentation effects and component-level quality.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness)

General Terms

Design, Experimentation, Measurement

Keywords

Web search relevance, measurement, evaluation

1. INTRODUCTION

Traditional information retrieval (IR) evaluation methodologies (e.g., [2][6]) judge the relevance of the individual documents from a ranked list returned for a query, and compute a single performance score that is averaged across many queries. This method of assessing a search engine's result relevance manifests a high level of abstraction over the retrieval task by eliminating several sources of variability [5], enabling important experimentation, measurement, and evaluation efforts. However, this method ignores page elements other than the main document ranking, the surface-level representation of these elements, and holistic results-page characteristics such as coherence, diversity, and redundancy.

Search engine result pages (SERPs) play a critical role in the Web search process. Web searchers first interact with the SERP returned for their query and then with the retrieved results. On the SERP, each result has a summary and may include multimedia or links to additional documents. Around and interspersed within the ranked list are other page elements such as suggested spelling corrections, suggestions of follow-on queries, results from alternate queries, advertising links, and mini-rankings from other sources such as news, image, and video search.

Whole-page relevance (WPR) defines how well SERP components and the corresponding holistic attributes of the result page presentation respond to searchers' information needs. Despite its importance, whole-page relevance is seldom considered in IR evaluation. User studies (e.g., [4]), log analysis of user interaction with SERP components (e.g., [1]), and parallel A/B testing (e.g., [3]) can capture aspects of WPR but are limited in terms of factors such as scalability (user studies are costly and time consuming) or their ability to capture qualitative feedback (log analysis and A/B tests study only behaviors rather than users' rationales for them).

This poster presents an evaluation method for whole-page relevance. Our evaluation metaphor draws on teaching assessment and has judges consider the SERP responses to a query as though they were teachers grading school assignments from multiple students. While each student may have different styles and layout, overall they can be graded with respect to how well they address and satisfy the information needs represented in the assignment. Assignments can be graded both on component elements of their response (e.g., did they mention an important fact?) and on holistic aspects such as coherence, comprehension, and use of authoritative sources. This metaphor gives rise to our method's name: the *School Assignment Satisfaction Index* (SASI).

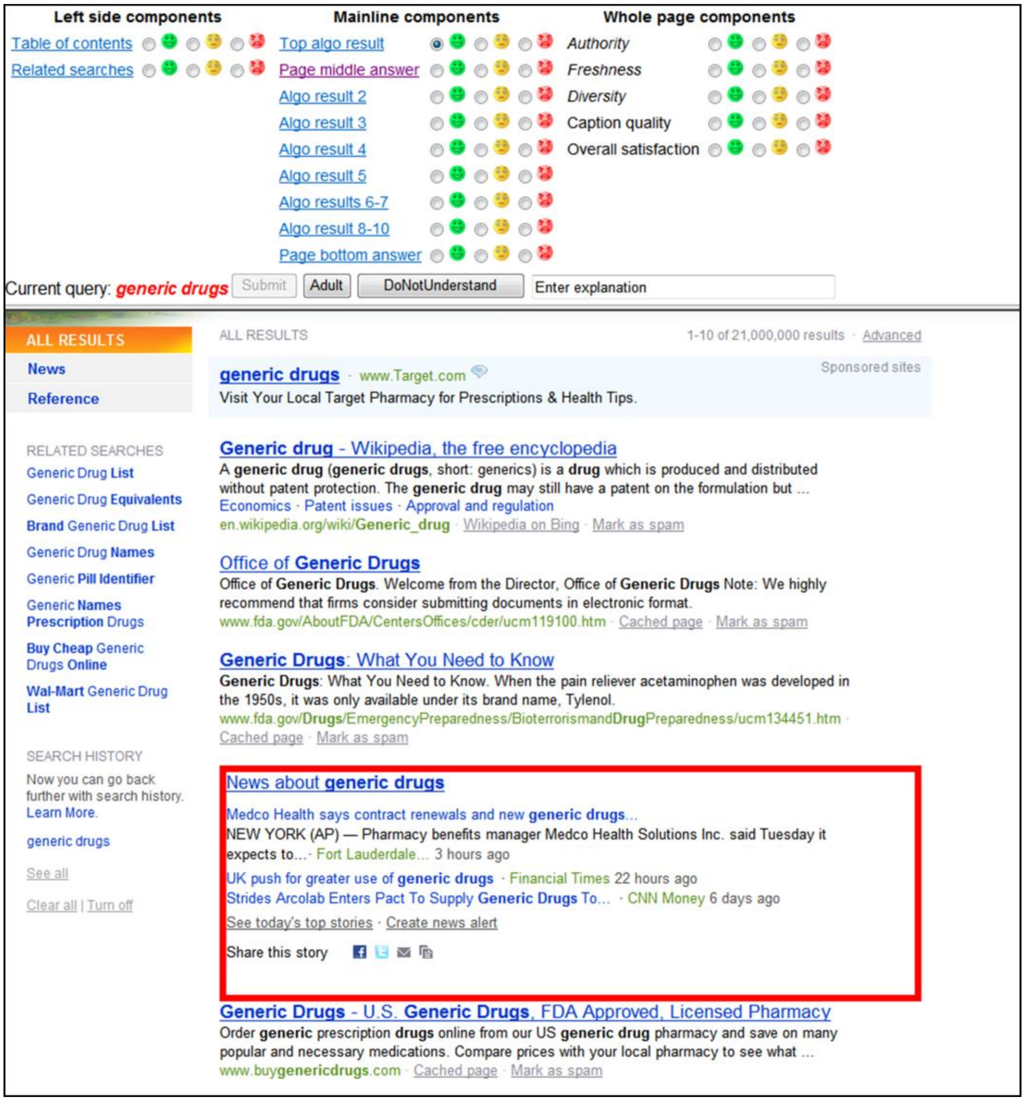
2. OVERVIEW OF THE SASI METHOD

The SASI method is a framework for judging aspects of SERP component and holistic relevance. It differs from traditional Cranfield-style evaluations [2] in three important ways: (i) judging the surface-level presentation on the SERP rather than the full content of documents, (ii) judging all components of the SERP rather than only the top-ranked algorithmic search results, and (iii) judging the SERP components in context rather than judging each document in isolation. Since SASI focuses on assessing the user's experience when interacting with the search engine, SERPs are captured and presented to judges in their entirety. However, there is no interaction with the interface components or any hyperlink clicking which may take the judge away from the SERP and potentially introduce bias from landing pages encountered. SASI focuses on judging only the surface-level representation of the page components, but an alternative WPR judging system could also judge the landing pages of all linked items.

3. EXAMPLE SASI JUDGE INTERFACE

Figure 1 shows our implementation of a SASI judging interface, with a judging pane and a pane showing a Microsoft Bing SERP for the query [generic drugs]. During judging, the interface can highlight the element currently being judged, which in Figure 1 is *page middle answer* – a universal search result in the middle of the page. In the figure, the first of the top-10 results (*top algo*) has already been rated as *good*. Note that this is just one example of a judging interface, and others could be used. For example, although we adopted an emoticon-based judging scale; numeric or label scales could also be used. The selection of a three-point scale is arbitrary; a two, four, or five point scale may be equally or more effective for providing SASI judgments.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.



Left side components		Mainline components		Whole page components	
Table of contents	○ ○ ○	Top algo result	● ○ ○	*Authority*	○ ○ ○
Related searches	○ ○ ○	Page middle answer	○ ○ ○	*Freshness*	○ ○ ○
		Algo result 2	○ ○ ○	*Diversity*	○ ○ ○
		Algo result 3	○ ○ ○	Caption quality	○ ○ ○
		Algo result 4	○ ○ ○	Overall satisfaction	○ ○ ○
		Algo result 5	○ ○ ○		
		Algo results 6-7	○ ○ ○		
		Algo result 8-10	○ ○ ○		
		Page bottom answer	○ ○ ○		

Current query: *generic drugs*

ALL RESULTS 1-10 of 21,000,000 results · Advanced

Sponsored sites

generic drugs · www.Target.com
Visit Your Local Target Pharmacy for Prescriptions & Health Tips.

RELATED SEARCHES
Generic Drug **List**
Generic Drug **Equivalents**
Brand Generic Drug **List**
Generic Drug **Names**
Generic **Pill Identifier**
Generic **Names Prescription** Drugs
Buy Cheap Generic Drugs **Online**
Wal-Mart Generic Drug **List**

SEARCH HISTORY
Now you can go back further with search history. Learn More.
generic drugs
See all
Clear all | Turn off

Generic drug - Wikipedia, the free encyclopedia
A **generic drug** (**generic drugs**, short: generics) is a **drug** which is produced and distributed without patent protection. The **generic drug** may still have a patent on the formulation but ...
Economics · Patent issues · Approval and regulation
en.wikipedia.org/wiki/**Generic_drug** · Wikipedia on Bing · Mark as spam

Office of **Generic Drugs**
Office of **Generic Drugs**. Welcome from the Director, Office of **Generic Drugs** Note: We highly recommend that firms consider submitting documents in electronic format.
www.fda.gov/AboutFDA/CentersOffices/cder/ucm119100.htm · Cached page · Mark as spam

Generic Drugs: What You Need to Know
Generic Drugs: What You Need to Know. When the pain reliever acetaminophen was developed in the 1950s, it was only available under its brand name, Tylenol.
www.fda.gov/**Drugs**/EmergencyPreparedness/BioterrorismandDrugPreparedness/ucm134451.htm · Cached page · Mark as spam

News about **generic drugs**
Medco Health says contract renewals and new **generic drugs**...
NEW YORK (AP) — Pharmacy benefits manager Medco Health Solutions Inc. said Tuesday it expects to... · Fort Lauderdale... 3 hours ago
UK push for greater use of **generic drugs** · Financial Times 22 hours ago
Strides Arcolab Enters Pact To Supply **Generic Drugs** To... · CNN Money 6 days ago
See today's top stories · Create news alert
Share this story

Generic Drugs - U.S. **Generic Drugs**, FDA Approved, Licensed Pharmacy
Order **generic** prescription **drugs** online from our US **generic drug** pharmacy and save on many popular and necessary medications. Compare prices with your local pharmacy to see what ...
www.buy**genericdrugs**.com · Cached page · Mark as spam

Figure 1. SASI judging interface example. *Page middle answer* is currently being judged.

4. CONCLUSIONS

We have introduced SASI, a new evaluation method that deliberately considers only surface-level information on a search results page, enabling evaluation of whole-page relevance. While sharing some similarities with traditional search evaluation, the SASI method does not generate a reusable test collection because the judgments are not independent of other components shown on the page. Instead it gives visibility into different search system characteristics, that are normally evaluated using user studies or A/B testing. SASI is an efficient form of experiment; initial testing with the judge interface described herein revealed that a whole SERP can be judged in the time it takes to judge two documents in Cranfield-style experiments. We have also performed further investigations of the utility of SASI for evaluating whole-page Web search relevance, revealing that it provides insights on likely user perceptions of relevance that are unavailable via traditional IR evaluation methods (e.g., component-level quality and search engine branding effects). In targeting whole-page relevance, SASI provides a useful complement to document relevance approaches for assessing search performance. In future work we will refine SASI to further our understanding of whole-page relevance.

REFERENCES

[1] Clarke, C.L.A., Agichtein, E., Dumais, S., and White, R.W. (2007). The influence of caption features on clickthrough patterns in web search. *Proc. SIGIR*, 135-142.

[2] Cleverdon, C.W. (1960). ASLIB Cranfield research project on the comparative efficiency of indexing systems. *ASLIB Proceedings*, XII, 421-431.

[3] Kohavi, R., Henne, R., and Sommerfeld, D. (2007). Practical guide to controlled experiments on the web: listen to your customers not to the HiPPO. *Proc. SIGKDD*, 959-967.

[4] Turpin, A., Scholer, F., Järvelin, K., Wu, M., and Culpepper, J.S. (2009). Including summaries in system evaluation. *Proc. SIGIR*, 508-515.

[5] Voorhees, E.M. (2008). On test collections for adaptive information retrieval. *Information Processing and Management*, 44(6): 1879-1885.

[6] Voorhees, E.M. and Harman, D. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.