Relevance Criteria for E-Commerce: A Crowdsourcing-based Experimental Analysis

Omar Alonso
A9.com
Palo Alto, CA (USA)
oralonso@gmail.com

Stefano Mizzaro
Dept. of Mathematics and Computer Science
University of Udine
Udine (Italy)
mizzaro@dimi.uniud.it

ABSTRACT

We discuss the concept of relevance criteria in the context of e-Commerce search. A vast body of research literature describes the beyond-topical criteria used to determine the relevance of the document to the need. We argue that in an e-Commerce scenario there are some differences, and novel and different criteria can be used to determine relevance. We experimentally validate this hypothesis by means of Amazon Mechanical Turk using a crowdsourcing approach.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software — performance evaluation

General Terms

Measurement, performance, experimentation

Keywords

IR evaluation, relevance, relevance criteria, user study

1. INTRODUCTION

Relevance is a key notion for information retrieval. It has been extensively studied in the past, as surveyed in [4], and it is still under investigation, see, e.g., [5, 6]. One important research issue is the set of *relevance criteria*, i.e., the attributes that determine whether an item is relevant or not. Topicality is the classical relevance criterion; additional criteria have been studied extensively in the 90s, and the main outcomes are summarized by Barry and Schamber in [2].

Electronic Commerce, commonly known as e-Commerce, consists of the buying and selling of products or services over the Internet. A large fraction of e-Commerce is performed in customer facing Websites where search is the predominant activity. After all, if the customers can't find an item to buy they are likely to go to a different e-Commerce site.

How can one frame the concept of relevance in the context of e-Commerce? More specifically, are there any differences in beyond topical criteria for e-Commerce search? This question makes sense when one adopts a historical perspective. After the work by Barry and Schamber in the 90s, since the intersection among the sets of elicited relevance criteria was high, it seemed reasonable to assume that all — or almost all — criteria had been found [2]. This view has not changed for 10 years, but if we look at Barry and Schamber criteria from the e-Commerce viewpoint we see that some of the criteria do not apply and that perhaps new ones are needed. After all, the work on relevance criteria has been done in a pre-eCommerce era, and most of those criteria were based on finding documents, not on *buying* a product.

We propose the following nine e-Commerce specific relevance criteria, derived from some scenarios not shown for space limitations: brand name, product name, price/value (cheap, affordable, expensive, not suspiciously cheap), availability, ratings & user reviews, latest model/version, personal aspects, perceived value, and genre & age.

2. EXPERIMENT

We ran some experiments to find out if the above e-Commerce oriented criteria are perceived as more important for e-Commerce needs than for classical information needs, and vice versa if classical criteria are perceived as more important for classical needs than for e-Commerce needs.

We used for our evaluation study Amazon Mechanical Turk (MTurk, `http://www.mturk.com/`), a crowdsourcing platform. Very recently, crowdsourcing has emerged as a viable alternative to conduct large scale evaluation of different types of experiments for a wide range of applications like relevance evaluation [1] and user studies [3]. This recent research has shown that MTurk results are reliable and very useful for gathering extra feedback, as we will see below.

We asked the workers to rate the relevance criteria for some induced and realistic needs. We used 83 needs, derived by monitoring for two weeks the most frequent queries in Yahoo! Buzz (`http://buzzlog.buzz.yahoo.com`) and top product searches in Amazon (`http://a9.com`). 35 of them are about e-Commerce such as: `wii fit`, `nintendo ds`, `xbox 360`. And 48 are classical informational needs such as: `IRS tax forms`, `government jobs`, `health insurance`.

For each need, we listed 17 criteria: nine concern e-Commerce (those presented in Sect. 1), and eight are selected among the classical ones by Barry and Schamber [2] (accuracy & validity, consensus within the field, content novelty, depth & scope, presentation, recency, reliability, and verifiability).

In the HITs (Human Intelligence Task), e-Commerce and non e-Commerce needs criteria were mixed and presented alphabetically. The workers were not told whether they were concerned with e-Commerce or classical search. The task consisted in selecting the criteria (one or more) for a given query. We paid each worker $0.01 cents (US). We planned

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

		Criteria	
		e-Commerce	Non e-Commerce
Needs	e-Commerce	**659** (**26.90**%)	516 (21.06%)
	Non e-Commerce	539 (22.00%)	**736** (**30.04**%)

Table 1: Main results.

7 workers for each need (query), for a total of 83 HITs (581 assignments). We got 100% of the answers after two days.

3. RESULTS AND CONCLUSIONS

The workers performed 2450 criteria selection in total (4.2 on average for each assignment), almost equally distributed between e-Commerce and non e-Commerce needs (1175 vs. 1275) and between e-Commerce and non e-Commerce criteria (1198 vs. 1252). Criteria distribution is rather uniform: Accuracy was selected 309 times, Availability 264, and any other criteria between 162 and 81 times.

Turning to the hypothesis being tested, the results shown in Tab. 1 quite confirm the hypothesis, as they show that indeed there is a good correspondence: a total of 56.94% criteria are classified according to the need kind (i.e., either an e-Commerce criterion was selected for an e-Commerce need or a non e-Commerce criterion was selected for a non e-Commerce need), and 43.06% are classified in the opposite way (the two remaining cases). There is a particularly good match for non e-Commerce needs and criteria (30.04%).

This is further confirmed by a breakdown analysis on the single criteria, shown in Fig. 1. In the top chart (concerning 8 criteria that are not e-Commerce) all the criteria have been selected more in non e-Commerce queries than in e-Commerce ones (i.e., blue/darker bars are always higher than yellow/lighter bars). In the second chart (concerning 9 criteria that are e-Commerce), the converse is usually true, but not for all criteria (genre and personal aspects are the exceptions). Accuracy and depth score high for non e-Commerce (also consistent with [2]). The availability, price, and brand name of a product are the driving criteria for e-Commerce. It can be argued that genre and personal aspects are less e-Commerce than the other ones. However, maybe some of our instructions were not very clear for the workers, so further experimentation needs to be done.

As part of the experiment, workers have the option to leave comments about their task. This was very useful for gathering extra data points that were not covered in the criteria presented. We list here some of the most interesting comments, collected from both this experiment and a previous one on a reduced data set, that suggest new criteria or explicitly confirm the classical ones (emphasis added):

- About buying boots: "When looking for a product online there is also the need to see *return policy* information as well *aid to resolve problems* should they arise."
- About buying a black polo shirt: "I do not worry about to many things when buying shirts. My primary focus would be on the *price and value*."
- About buying an engagement ring: "It would be important for me to know the source is valid and *not a spam of phishing attempt*."
- About buying 2009 Hyundai Genesis: "I *trust* Hyundai."

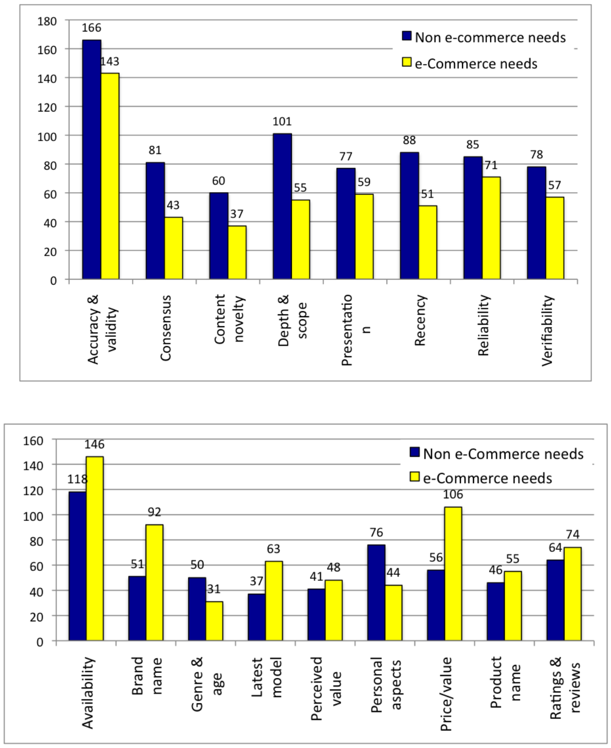

Figure 1: Criteria selection.

- About weight gain: "I would also want what I find to be *consistent with what I already know* (although, preferably, also expanding upon my existing knowledge base)."
- About Dakota Fanning: "I look for *reliable*, *current*, and *accurate* content."

In this short paper we have analyzed the notion of relevance in the context of e-Commerce. We presented evidence that when users are interested in buying products, they apply different criteria when deciding relevance. Future work includes running similar experiments with larger data sets.

4. REFERENCES

[1] O. Alonso, D. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[2] C. L. Barry and L. Schamber. Users' criteria for relevance evaluation: A cross-situational comparison. *IP&M*, 34(2/3):219–236, 1998.

[3] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *CHI '08: Proceeding of the 26th SIGCHI*, pages 453–456, 2008.

[4] S. Mizzaro. Relevance: The whole history. *JASIS*, 48(9):810–832, 1997.

[5] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *JASIST*, 58(13):1915–1933, 2007.

[6] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: behavior and effects of relevance. *JASIST*, 58(13):2126–2144, 2007.