The economics of database searching

August 11 2011

(PhysOrg.com) -- Searching the internet might seem simple, but applying a little bit of economic theory to information retrieval can shed some light on the best search strategies to adopt, according to researchers.

Dr. Leif Azzopardi from the School of Computing Science at the University of Glasgow took production theory from microeconomics and applied it to the process of searching the internet, or any other database system.

In economics, production theory deals with the process of turning inputs, like capital and labour, into outputs, like goods and services. Production theory helps to determine how resources can be efficiently turned into such products i.e. least effort, most gain.

Working on the premise that searching for information requires effort, such as assessing documents, and submitting queries, Dr. Azzopardi sought to consider what search strategies a user should employ to efficiently undertake a search when looking for a number of relevant documents.

Dr. Azzopardi said: “Short queries can be quite effective for finding one highly-relevant document, but searching for a number of relevant documents often requires numerous queries to be posed. Generally speaking, a user will only examine the first page or so of the result list.”

Dr. Azzopardi said: “We also know that longer queries are more

effective, and that there is often more relevant items on subsequent pages. So are people being lazy when they search or are they being strategic and rational when they search?"

"Being able to answer such questions is important for interactive information retrieval because while behavioural and observational studies have been conducted, there is a lack of formal theory to explain why such observations are witnessed."

Dr. Azzopardi conducted a simulated analysis that varied the way in which a simulated user interacted with three different types of database search methods: BM_{25}, Boolean and TFIDF.

Boolean searches use logical connections to search terms – using AND, OR, NOT; and are typically used in patent or library search systems.

BM25 ranks documents based on relevance using the incidence of key words searched that are contained in the document; and TFIDF (term frequency-inverse document frequency) is another method of searching for relevant keywords within a document.

By applying production theory from economics Dr. Azzopardi was able to identify which search strategies users of different retrieval systems should use.

Overall, he found that BM25 systems supported a greater variety of search strategies than Boolean or TFIDF. However, the most 'cost-efficient' search on a BM25 system involved examining only the first page or so of results and then posting further queries until the desired level of gain was achieved. This finding is consistent with how users search the internet using commercial search engines.

On the other hand using Boolean search systems suggests users would

have to delve deeper into the result listings looking at 100s of documents per query and issuing substantially more queries to achieve the same level of gain. However, this finding is consistent with how patent searchers interact with Boolean based patent systems.

Dr. Azzopardi said: "This work provides the foundations on which to build formal methods for describing, understanding and explaining the interactions between a user and system."

"It also shows that we can apply economics to human computer interaction, more generally, and therefore we can predict how a user will utilize a system."

More information: The research was presented at a conference of the Special Interest Group on Information Retrieval in Beijing in July and is published on the website of the Association for Computer Machinery. (portal.acm.org/citation.cfm?id=2009923)

Provided by University of Glasgow