Decision Support Systems 54 (2013) 904–918

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

# Concept comparison engines: A new frontier of search

Alan S. Abrahams [a],*, Reza Barkhi [b]

[a] Department of Business Information Technology, Pamplin College of Business, 1007 Pamplin Hall, Virginia Tech, Blacksburg, VA 24061, USA
[b] Department of Accounting and Information Systems, Pamplin College of Business, 3007 Pamplin Hall, Virginia Tech, Blacksburg, VA 24061, USA

ARTICLE INFO

*Article history:*
Received 11 February 2011
Received in revised form 29 June 2012
Accepted 20 September 2012
Available online 27 September 2012

*Keywords:*
Preferential choice
Search engine
Text mining

ABSTRACT

In a traditional search engine interaction scenario, a user begins with a certain concept and finds documents that are similar to their concept. However, the user may wish to compare alternatives and a search capability should compare concepts and present the best alternatives. This task can be difficult without proper decision aids. We propose a concept comparison engine as a decision support tool that may be used to compare attributes of different alternatives and aid in making an informed selection. We describe an architecture and an interaction scenario and implement a prototype. We propose a number of evaluation metrics for measuring the viability of different terms for the purpose of comparing concepts. In scripted experiments, orderings for candidate terms from the prototype are compared to gold standard ranking lists from structured external sources. Our results indicate that a Rankor analysis may be promising as a measure of the differentiating power of candidate terms a user might choose to support concept comparison.

## 1. Introduction

The decision making process often involves defining the problem, gathering the data, building a model, generating alternatives, and selecting a good alternative [24]. However, with the overwhelming volume of free-form textual information available on the internet, it is extremely difficult to process the avalanche of results from a single search, let alone compare alternatives. Research on the use of decision support systems for preferential choice tasks has found that users seek to minimize their overall expenditure of effort, even if this negatively impacts decision quality [49]. Users frequently stop early before a comprehensive search and analysis is complete [9]. Furthermore, users quickly become overwhelmed by the number of available alternatives, unless the large number of alternatives can be easily accessed by attribute values [28].

Modern search engines are great at giving relevant information on one topic, but they are not properly designed for comparing among alternatives. For example, while Google and Yahoo are excellent for finding relevant documents on the state of Virginia, they are not particularly suited for comparing Virginia to Alaska. Current search engines give the user some information on the target item: for instance, links to the state's home page, tourism pages, and wiki pages. However, often in human decision making, the merit of a solution can be understood only when it is compared against other alternatives. For example, while someone may say that Virginia is a great place to do business, the question remains, compared to what? Also, on what basis are you comparing the states—population, climate, transportation, natural reserves like coal and oil? In this paper, we propose a framework, which attempts to shift the search engine decision support task from solely presentation of relevant information on a single topic to providing a decision aid for comparison of multiple topics.

Consider MountaineerGear, a (fictitious) manufacturer of mountain climbing gear, looking to expand its network of direct-to-consumer outlet stores. MountaineerGear uses a traditional search engine to find some subjective information from popular journalistic sources (e.g. articles on "the best places to go mountain climbing"). However, this journalistic information is imperfect: it may be influenced by editorial biases (pandering to paying advertisers), or may simply be hamstrung by each journalist's limited expertise. After some research, MountaineerGear determines that quantitative information about mountains and mountain climbing is available from governmental and commercial data sources. For example, government data providers like the United States Geological Survey can provide the locations of the 100 highest peaks in the country, but this tells MountaineerGear where mountains are, not where people like to go to climb mountains. Commercial data providers can provide the locations of all competitor stores that are listed under the industry classification NAICS category 33999 ("Climbing and Rappelling Gear"). This would show where some mountain-climbing-related stores are currently operating, but not where are the good future locations. For an in-depth comparison MountaineerGear might consider using a search engine to gather documents from each candidate location. But, each candidate location search uncovers dozens of documents. MountaineerGear must synthesize the information from the document collection for each candidate location, and then attempt to compare locations against each other. The process is arduous and likely to be biased by eventual search

* Corresponding author. Tel.: +1 540 231 5887; fax: +1 540 231 3752.
*E-mail addresses:* abra@vt.edu ,abrahamsalan@yahoo.com (A.S. Abrahams), reza@vt.edu (R. Barkhi).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2012.09.014

fatigue. Inadequate review of locations could lead to high losses or high opportunity costs. A concept comparison engine (CCE) would allow MountaineerGear to compare candidate locations. We propose that an automated comparison of locations, showing the extent to which web pages in each location refer to "mountain climbing," could provide a rapid and informative "heat-map" of mountain climbing enthusiasts. Such a comparison is outside the gamut of traditional decision support tools, but would be useful, and complementary to the above data sources.

In Table 1, we suggest a number of sample scenarios where a CCE may be helpful. In each case, the prospective user and their goal are specified. We describe this as a two-step process. In Step 1, we describe what alternatives the user would compare (i.e. the concept list), and what attributes the user could use to compare these alternatives, using the CCE to analyze both structured (Step 2a) and unstructured (Step 2b) data sources. For instance, in the MountaineerGear example (Example 1 in Table 1), the marketing manager uploads a list of US cities (available from the United States Geological Survey) and compares them by both population (Step 2a), and the relative term frequency for "mountain climbing" in each city (Step 2b). In Example 2 in Table 1, a high school graduate may wish to find an affordable college with a strong business ethics program. The graduate would upload a list of AACSB accredited schools to the CCE (Step 1) which may be part of a spreadsheet from the AACSB which has tuition fees for each college (Step 2a). The CCE would then gather a document collection for each college on the AACSB list, and allow the graduate to find the term frequency for "business ethics" for each candidate college (Step 2b). The mash-up of structured tuition fee data (Step 2a) and unstructured program content information (Step 2b) helps the graduate compare colleges. Finally, in Example 6 in Table 1, a College Career Services Director for University X may be looking to build recruitment relationships with the best companies. The director may use a CCE to choose companies within a 150 mile radius that have the high revenues or profits, and have alumni amongst their listed employees (i.e. employees that specify "University X" on their web or social media pages). More detailed information about the CCE implementation, including a full worked example (Example 4 in Table 1) listing concepts to compare, evaluation criteria, data collection from structured and unstructured sources, and comparing the alternatives, is provided in the Online Supplement to this paper.

The CCE is intended to provide a layer of additional functionality on top of traditional search, to reduce total expenditure of effort in preferential choice tasks and improve decision quality [49]. The CCE is also intended to allow alternatives described in unstructured text to be more easily accessed by attribute, potentially improving perceived usefulness of the DSS [28]. In short, the CCE is aimed at easing the process of not just gathering information (as in traditional search), but also of comparing alternatives.

In light of the need for mechanisms to automatically compare concepts that are described in both free-form text documents and in tabular data sources, we make a number of contributions. Firstly, we define a model for concept comparison from text documents, and describe an architecture and implementation for this new type of DSS. Secondly, we propose a number of CCE evaluation metrics, and evaluate the applications and limitations of the CCE through various experiments.

Section 2 briefly describes the related work in the information retrieval and information extraction fields. Section 3 illustrates the proposed CCE model. Section 4 describes and illustrates an interactive decision support session where the user works with the CCE on a "choice-between-alternatives" decision making task. Section 5 shows the general architecture of a CCE and Section 6 remarks on our specific implementation. Section 7 specifies metrics for determining the plausibility and validity of the comparisons produced by the CCE. Section 8 reports the results of a large number of experiments with the CCE system and explains the implications of this experience for how CCE's should be built, configured, and used.

## 2. Related work

Information retrieval (IR) on the web, commonly known as "web search," is the process of gathering documents relevant to a particular concept, and ranking them by relevance [50]. The output is a ranked list of search results (pointers to relevant documents). A search on "Boston Population," for instance, may yield millions of documents relevant to Boston and its population. In information extraction (IE) [13,14,19,29], relevant facts are extracted from available documents, to fill in the blanks in a user's knowledge. For example, the query "Boston population" might result in the fact "Boston population is 589,141." We distinguish concept comparison (CC) as a value-added feature on top of retrieval and extraction. Comparing Boston and Philadelphia, by population, for instance, requires first retrieval of relevant documents (discussing the population of both cities), and then extraction of the pertinent facts, before comparison can be operationalized. The state of the art in information retrieval

**Table 1**
Example CCE applications.

| Example number | User | Goal | Structured data source | Step 1: List alternatives | Step 2a: Acquire attributes from structured sources | Step 2b: Compute aggregates from unstructured text |
|---|---|---|---|---|---|---|
| 1 | Manufacturer of mountain gear | … wants to find good locations for outlet stores | United States Geological Survey (USGS) | List of US cities | Population (high is good) | Term frequency for "mountain climbing" in document collection for each city |
| 2 | High school graduate | … wants to find an affordable college with a strong business ethics program | Association to Advance Collegiate Schools of Business (AACSB) | List of AACSB-accredited business schools | Annual tuition | Term frequency for "business ethics" in document collection for each business school |
| 3 | Parent | … wants to find schools with a low student-teacher ratio and focus on graduates going to Ivy-league colleges | National Center for Education Statistics Elementary/ Secondary Information System (NCES ELis) | List of secondary schools | Pupil/teacher ratio | Term frequency for "ivy league" in document collection for each school |
| 4 | Marketing manager at chemical manufacturer | … wants to find candidate locations for new, toxic product for the coal mining industry | United States Geological Survey (USGS) | List of US Cities | Population (low is good) | Term frequency for "coal mining" in document collection for each city |
| 5 | Manufacturer of fishing rods/fishing enthusiast | … wants to find places where freshwater fishing is feasible and popular | USGS National Water Information System (NWIS) | List of US water sites | Water quality metrics | Term frequency for "fishing" in document collection for each site |
| 6 | College career services director for University X | … wants to find employers for graduating seniors | Securities and Exchange Commission (SEC), Forbes 500 list | List of public and private co.'s | Revenues, Profits, Locations | Term frequency for "University X" in document collection for each company |