Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 169 – 178

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Evaluating the suitability of Web search engines as proxies for knowledge discovery from the Web

Laura Martínez-Sanahuja*, David Sánchez

UNESCO Chair in Data Privacy, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili Av.Països Catalans, 26, 43007 Tarragona, Catalonia, Spain

Abstract

Many researchers use the Web search engines' hit count as an estimator of the Web information distribution in a variety of knowledge-based (linguistic) tasks. Even though many studies have been conducted on the retrieval effectiveness of Web search engines for Web users, few of them have evaluated them as research tools. In this study we analyse the currently available search engines and evaluate the suitability and accuracy of the hit counts they provide as estimators of the frequency/probability of textual entities. From the results of this study, we identify the search engines best suited to be used in linguistic research.

Peer-review under responsibility of KES International

Keywords: Web search engines; hit count; information distribution; knowledge discovery; semantic similarity; expert systems.

1. Introduction

Expert and knowledge-based systems rely on data to build the knowledge models required to perform inferences or answer questions. Yet, their performance is tied to the availability and coverage of the electronic data they use as knowledge sources. In this respect, the success of the Internet has multiplied the amount of electronic resources that are freely available for research. By exploiting these large data sources, it has been possible to build expert systems with a performance we were only able to imagine few years ago. For instance, the well-known Watson expert system by IBM was able to win the *Jeopardy!* quiz show in front of expert human players by exploiting more than

* Corresponding author. Tel.: +34 977559657; fax: +34 977559710.
E-mail address: laura.martinez@urv.cat (L. Martínez-Sanahuja), david.sanchez@urv.cat (D. Sánchez)

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.123

200 million pages of linguistic electronic content, which included the whole Wikipedia [1] .

Many expert systems rely on linguistic data for knowledge discovery, especially those dealing with textual inputs/outputs. Indeed, most of the information that is being produced nowadays is textual, because it constitutes the natural mean of interaction among human actors. In this respect the World Wide Web is currently the largest freely available source of electronic data, most of which is of linguistic nature. As we realize more and more on the importance of the availability of *big* linguistic data for the development of expert systems, the more appealing the use of the Web as knowledge source it becomes. In fact, the Web is so large, heterogeneous and up-to-date that it is said to be a faithful representation of the current information distribution at a social scale [2], an argument that has been supported by recent works [3-5], which considered the Web as a realistic proxy for social knowledge.

Because of its interesting features, many researchers have used the Web as a knowledge source and, more specifically, to estimate the distribution of linguistic data from the frequency/probability of (co-)occurrence of entities of interest (e.g., textual terms, concepts, bigrams, etc.). In this way, researchers attempt to alleviate the constraints imposed by static linguistic corpora that, despite being reliable and unambiguous, are limited in terms of size, coverage and updates, thus usually producing data sparseness problems [5].

The usual low-entry-cost way to access to Web data is via a commercial Web search engine (WSE). Indeed, frequencies or probabilities for some phenomenon of interest can be straightforwardly estimated from the hit count provided in the search engine's result page [6]. Early works using hit counts tackled the identification of translation for compositional phrases [7], the discovery of synonyms [8] or the assessment of frequencies of bigrams [9]. More recent works include building models of noun compound bracketing [10], automatic ontology learning [4, 11-13], large-scale information extraction [14, 15], semantic similarity estimation [2, 16, 17], topic discovery [18], user profiling [19] or disclosure risk assessment in textual documents [3, 5, 20].

When WSEs are used as proxies for the Web information distribution in tasks such as the former ones, the outcomes of these tasks closely depend on the suitability of the hits counts as frequency estimators. Yet, many researchers relegate the choice of the search engine, and employ the WSE they are familiarized with, thus potentially compromising their research results; in fact, the search engines most commonly used in research are also those most used in general: Google, Bing and Yahoo! [2, 5, 17, 21]. Some studies have criticized these choices and questioned the usefulness of well-known WSEs as research tools due to the issues they present (ambiguity, constrained query languages, commercial bias, arbitrariness of hit counts, etc.) [22].

The study we conduct in this paper aims at bringing some light into these issues and, specifically, to the following questions: *are WSE really effective as proxies for Web information distribution?*, *how far the choice of a particular WSE may influence the outcomes of the task to which it is applied?*, and ultimately, *which is the WSE best suited for linguistic research?* For this purpose, we survey and systematically analyse most of the WSEs currently available (being commercial or not) and select those able to provide hit counts that can be used as general-purpose estimators of term frequencies. The selected search engines are evaluated from both qualitative and quantitative perspectives. In the former case, we define a set of quality criteria that the WSE should fulfil in order to be considered an appropriate research tool (i.e., mathematical coherence of hit counts, flexibility of the query language, non-exact search capabilities and access restrictions). In the latter application-oriented evaluation, we use the hit count of the different WSEs in one of the central tasks of computational linguistics (i.e., the estimation of the semantic similarity between concepts), and objectively measure the accuracy of the outcomes they provide. From the results of these evaluations, we identify the search engines(s) best suited for linguistic research.

Many studies have evaluated the effectiveness of WSEs as information retrieval tools [23-27]. However, few works have analysed the suitability of WSEs' hit counts for linguistic research. In [28], the author measured the correlation between the hits counts provided by several well-known search engines for as set of queries, whereas in [29], the coherence of the hit counts was tested against the actual number of web sites indexed by the WSE; finally, the studies performed in [30, 31] focused on analysing the reliability of the hit counts through time. All the former works focused on the three most used WSEs: Google, Yahoo and Bing/Live Search. As main contributions over these works, our study provides a more up-to-date survey, considers a much broader spectrum of WSE (going beyond the "usual suspects" Google, Bing and Yahoo) and, in addition to the application-agnostic analysis of WSEs' hit counts, we provide an application-oriented evaluation in a core task of linguistic research: the estimation of the semantic similarity between textual terms; with this we aim not only at assessing the potential of WSEs' hit counts, but also to measure their actual performance in a realistic research setting.