Search as research practices on the web: The SaR-Web platform for cross-language engine results analysis

Davide Taibi[1], Richard Rogers[2], Ivana Marenzi[3], Wolfgang Nejdl[3], Qazi Asim Ijaz Ahmad[3], Giovanni Fulantelli[1]

[1]Consiglio Nazionale delle Ricerche, Istituto per le Tecnologie Didattiche, Palermo
{davide.taibi, giovanni.fulantelli}@itd.cnr.it

[2]Media Studies, University of Amsterdam
rogers@uva.nl

[3]L3S Research Center, Hannover
{marenzi, nejdl}@L3S.de

ABSTRACT
Search engines are the most utilized tools to access information on the Web. The success of large companies such as Google owes to their capacity to conduct users through the vast troves of knowledge and information online. Recently, the concept of *search as research* has been used to shift the research focus from workings of information-seeking tools towards methods for the social study of Web and particularly the social meanings of engine results. In this paper, we present SaR-Web, a web search tool that provides an automatic means to carry out *search as research* on the Web. It compares the results of same (translated) queries across search engine language domains, thereby enabling cross-linguistic and cross-cultural comparisons of results. SaR-Web outputs enable the comparative study of cultural mores as well as societal associations and concerns, interpreted through search engine results.

CCS Concepts
Information systems~Web searching and information discovery

Keywords
Digital methods, Search as research, cross-language analysis.

1. SEARCH AS RESEARCH
The World Wide Web is now widely used in most populated places throughout the world and has become the most popular gateway to find most types of information. Although English for a long time had been considered the lingua franca of the Internet [4], the Web has internationalised. The multiplicity of languages and (national) cultures, online, are aspects to be taken into account when searching for content on the Web but also when developing research strategies for studying Web data. As the amount of content increases and is searchable in multiple languages, new opportunities emerge for the study of the interplay of language and content. As a case in point, specific studies on Wikipedia have pointed out that each language edition contains its own cultural viewpoints on a large number of concepts [3][5].

WebSci '16, May 22-25, 2016, Hannover, Germany
ACM 978-1-4503-4208-7/16/05.
http://dx.doi.org/10.1145/2908131.2908201

Furthermore, the local-domain versions of search engines (e.g. Google or Bing .de, .it, .co.uk and so on) in different countries return particular results that are aimed to meet the expectations of people in those specific countries, based on previous users' behavior (e.g. what sources users have clicked in previous searches, inlinks received by sites, and users´ click count and freshness). Reflecting on the qualitative, epistemological value of the information that can be obtained through algorithmic search from the social point of view has caught the attention of media researchers and sociologists. The web can be considered as an important site to study everyday life, including cultural mores as well as societal positions and concerns. In this context, the concept of *search as research* has been coined to shift the research focus from the mechanics of information-seeking tools ('search research') towards formulating specified and underspecified queries and making social research findings with engine outputs [6]. In the book *Digital Methods* Rogers presents a methodological outlook for research with the web that aims to answer broader social and cultural research questions rather than focusing on the medium itself or solely on online culture. The proposition is to think along with how search engines and platform handle natively digital objects, and redeploy the methods as well as the outputs. Hyperlinks, hits, likes, tags, date stamps and other natively digital objects may be used to study the medium (for example, in order to improve the search results), but they also may show instead the "politics of association" (who links to whom) as well as "politics of memory" (where Wikipedia language editions of the 'same' historical event have distinctively different 'facts'). As social research shifts from being about the web (e.g., the digital divide or how much of society and culture is online), and moves to the opportunities web data afford, the research agenda begins to shift towards creating methods and techniques to study cultural change and societal condition with the web, also known as 'digital methods' ([7], p. 21). "Are engines placing alternative accounts of reality side by side, or do the results align with the official and the mainstream?" ([7], p. 31). When is the Web indifferent to the geographical location and language of its users, and when does it take into account the nations/countries as well as the language of the user? How can digital methods and social research take advantage of the linguistically and geographically grounded information?

2. THE SAR-WEB PLATFORM
The SaR-Web platform has been developed to support *Search as research* practices. It is based on a previous system called MWS-

Web (Multimodal Web Search), originally designed to facilitate the study of the web as a corpus [1]. The system provides a set of tools that facilitate the analysis of various types of web resources (websites, videos, images) and web genres (blogs, news, etc.) with a specific focus on the multimodal aspects [2]. Commercial search engines such as *Google, Yahoo and Bing* do not attempt to promote and encourage reflection on engine workings as well as outputs in any systematic way, e.g., by encouraging comparison and reflection on search results for the same query across different language domains. Such activities are valuable for reflecting upon and discovering new knowledge or information but also for undertaking *search as research*. For this reason we enhanced the functionalities of MWS-Web to support the investigation of broader research questions, described in this paper as SaR-Web.

The comparison of search results in different languages is often hindered by the difficulties in performing traditional textual analysis. To this end, in SaR-Web we implement a semantic based approach in which the comparison between search results in different languages is supported through visualization of semantic concepts, thereby overcoming the limit of textual descriptions.

SaR-Web provides word clouds in four languages (English, German, French and Italian) which highlight the most relevant keywords in localized Web sites. SaR-Web uses the Lucene API[1] to filter (using stop words), stem, index and search, and applies information retrieval techniques to text. The Lucene API is employed along with the Dandelion's Entity Extraction API[2] to generate (semantic) word clouds in the respective language. After a user searches for a keyword, the returned results (URLs) are obtained from the Bing search engine and are sent to the Dandelion's Entity Extraction API. The Entity Extraction API parses the content and sends back a response containing the extracted Wikipedia entities along with other information. The Wikipedia entities from the response are used to extract the DBpedia concepts (or keywords) that are in turn filtered and indexed for that specific search. After all the responses for that particular search are indexed, keywords along with their term frequencies are retrieved in order to create the word cloud.

In detail, the main tasks performed by SaR-Web are as follows:

1. *Localized search:* the keywords introduced by the user are searched by using the language and locale settings (e.g., "language:it loc:it"), so that only web pages from a specific country or region, and written in a specific language are returned.
2. *Named entity recognition*: the title and the snippet text from the body retrieved from the search engine results are elaborated with the Dandelion NER (Named Entity Recognition) service. This service returns the Wikipedia reference extracted by the NER procedure. This operation is performed for the four languages supported by SaR-Web.
3. *Semantic annotation*: SaR-Web transforms the Wikipedia reference in each language to the correspondent concept in the DBpedia knowledge base.
4. *Visualization*: the cloud is generated with the main concepts (or keywords) for the four languages (Figure 1).

2.1 Query: "Nuclear"

In this example we enter the query "Nuclear" (specified query) where we expect to obtain different results in different countries.

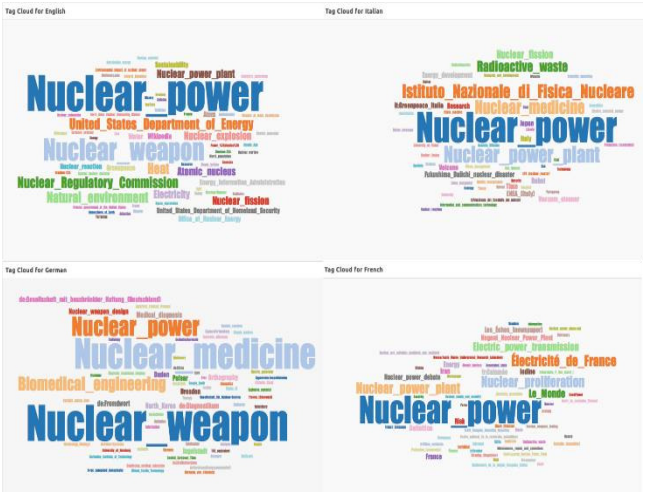

Figure 1: Concept Clouds for "Nuclear"[3]

The initial expectation is confirmed in that the concepts (or keywords) retrieved from the English, the Italian and the French results stress the nuclear energy while in the German cloud the most prominent concept, nuclear weapon, is associated to war.

3. FINAL REMARKS

Search as research is based on the fact that search engines are increasingly serving locally and linguistically grounded results and basing the results (among other signals) on crowd-sourcing, e.g. user clicks. By taking into account the users' behaviour, we have also explored the clouds produced by SAR-Web when considering only the first 10 results of the Bing search engine, which corresponds to the behavior of most Web users [10].

In particular, when considering the first 10 results, the query "nuclear" in the German local domain (.de) gave different results from those we obtained considering the first 20 results. In this case the most prominent concept associated to the keyword "nuclear" is Nuclear power also in German[4], providing a somewhat more positive sentiment as opposed to the clearly negative one (Nuclear_weapon) obtained before (Figure 1).

Similarly, SAR-Web produces different results when considering and processing the full text of the sources (top 10 websites) to simulate the behavior of a user who not only glances at the result list, but who investigates the results in more depth by looking at the content of the pages.

There are many other factors to be considered in Search as research studies. SAR-Web aims at supporting these studies through automated means, and reducing the need for manual intervention (e.g. to prepare the search engine, specify local-domain settings for the country specificity of the languages, rank lists of results and refine the query). However, moving from manual activities to automatic procedures is challenging, and finding solutions to address this fascinating and still open challenge requires the contribution of experts from different research fields and expertise such as computer scientists, sociologists and digital humanities experts.

[1] https://lucene.apache.org

[2] https://dandelion.eu/docs/api/datatxt/nex/v1/

[3] Hi-res version: http://www.pa.itd.cnr.it/websci16/nuclear.png

[4] Available online at: http://www.pa.itd.cnr.it/websci16/nuclear-10.png

4. ACKNOWLEDGMENTS

This work was partially funded by the European commission in the context of the ALEXANDRIA project (ERC advanced grant no: 339233)

5. REFERENCES

[1] Baldry, A.P. (2011a) *Multimodal Web Genres: Exploring Scientific English.* Como: IBIS.

[2] Baldry, A.P., Gaggia, A. and Porta, M. (2011) "Multimodal Web Concordancing and Annotation. An overview of the MCAWEB System", in Vasta, N., Riem N., A., Bortoluzzi M. and Saidero D. (eds.). Identities in Transition in the English-Speaking World, Udine: Forum Editrice Universitaria Udinese, pp. 39-60.

[3] Callahan, E.S. and Herring, S.C. (2011). Cultural bias in Wikipedia content on famous persons. Journal of the American Society for Information Science and Technology. 62, pp. 1899-1915

[4] Flammia, M. & Saunders, C. (2007). Language as power on the Internet. Journal of the American Society for Information Science and Technology, 58(12): 1899-1903.

[5] Pfeil, U., Zaphiris, P. and Ang, C.S. (2006). Cultural Differences in Collaborative Authoring of Wikipedia. Journal of Computer-Mediated Communication. 12, 1,88-113

[6] Rogers, R. (2013). Digital methods. Cambridge: MIT Press.

[7] Rogers R., Jansen F., Stevenson M. and Weltevrede E., "Mapping Democracy," Global Informaton Society Watch 2009, Association for Progressive Communications and Hivos, 2009, 47-57.

[8] Taibi D., Kantz D., and Fulantelli G. (2014). Supporting formative assessment in Content and Language Integrated Learning: the MWS-Web platform. International Journal Technology Enhanced Learning. 6, 4 (April 2014), 361-379. DOI=http://dx.doi.org/10.1504/IJTEL.2014.069042

[9] Zhang, J. & Lin, S. (2007). Multiple language supports in search engines. Online Information Review, 31(4), 516-532.

[10] Van Deursen, A. J., & Van Dijk, J. A. (2009). Using the Internet: Skill related problems in users' online behavior. Interacting with computers, 21(5), 393-402

.