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], Asim Ijaz[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. Most often search engine mechanisms based on ranking algorithms, indexes, and crawling strategies are hidden to users, but they are crucial for providing as well as studying search results. 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. *Search as research* practices are already applied in studies of both the live and archived Web, in order to critically study Web data for social research purposes, albeit with much manual effort. 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. Through making specified queries ("green tips") as well as varying degrees of underspecified queries ("nuclear" and "rights"), Sar-Web outputs enable the comparative study of cultural mores (each country's particular green tips) as well as societal associations and concerns (each country's ranked associations and concerns for nuclear and rights), interpreted through search engine results.

CCS Concepts

H.3.3 [Information Search & Retrieval]: Search Process;

H.1.2 [User-Machine Systems]: Human Factors

Keywords

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

1. INTRODUCTION

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 [13], the Web has internationalised. Nowadays both the sources as well as the gateways to them are available in various languages. 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. On the one hand, there is the question of the breadth and scope of technical as well as cultural content in each language online. Are there particular content divides that are language-based? On the other, one may make use of the kinds of content online in each language (and the absence of content), and thereby inquire into interests, priorities, biases and commitments that can be interpreted from the presence and absence of content per language domain. 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 [8][15]. Hecht and Gergle argue that the *world knowledge diversity* embedded in Wikipedia has important potentials for social research [10]. *Search as research* takes up both questions, though here we concentrate on the latter area of research, and extend its scope to the Web. How can the comparison of search engine results across language domains be made productive for social research?

Search engines are the most utilized tools to access information on the Web; most of them offer a variety of language aids and increasingly sophisticated tools to harvest information content from many languages, including multiple interface languages, translated search and cross-language search services [14][20]. Cross-Language Information Retrieval (CLIR) aims at facilitating information access across languages: it enables search using queries in one language to find content in another one [26][11]. Oard and Diekema in [27] identified three basic transformation approaches to CLIR: query translation, document translation, and interlingual techniques. A comparative study of online translation services for CLIR found that "although on average the machine translation (MT) approach usually provides a high quality translation for queries that consequently leads to high retrieval effectiveness in CLIR, the different MT systems can result in quite different retrieval behavior for individual queries for different language pairs and applications" [17].

Translating queries and results though is not enough to have a comprehensive overview of a topic on the Web, e.g., how it is

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

represented in search results, and what are its actual *value for* and *impact on* society.

Search engine ranking and crawling strategies are hidden to users, but they are crucial for providing and understanding search results. Where language is concerned, search engines use various techniques and analytics to obtain different rankings of search results across local-domain versions (e.g. Google or Bing .it, .de, co.uk and so on). The definition of "local" is broader than only based on country domains, though. In different countries the specific version of a search engine returns 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). The distinctiveness of the content in localized search engines can be casually confirmed by those who speak several languages and make the same (translated) query across multiple, local engines. This phenomenon was also confirmed by the study by the Digital Methods Initiative (DMI) [18] which compared the results of the query "Rights" in various languages in the local domains in Google in order to show differences in what matters to each language or culture. The study revealed hierarchies of rights type per country, where for example 'the right to roam' is particularly dear to Finland and the 'right to oblivion' or to forget to Italy, an issue subsequently taken up in European data privacy legislation concerning local domain Googles.

2. SEARCH AS RESEARCH

Reflecting on the qualitative, epistemological value of the information that can be obtained through algorithmic search, not only from the technical point of view (i.e., the study of the workings of Internet), but rather from the social point of view (the study of social and cultural with Web data), has caught the attention of media researchers and sociologists. The web is no longer to be considered only as a virtual realm, but rather increasingly 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 [16]. 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, aka 'digital methods' ([18], p. 21). "Are engines placing alternative accounts of reality side by side, or do the results align with the official and the mainstream?" ([18], 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?

Per geography, web results, then, may be viewed as hierarchies of (source) credibility, produced by engines' algorithmic directories, ranking systems, online metrics and optimization practices for particular places. The question is then is no longer to what extent the user concerns herself with the algorithmic techniques and impacts the frequent changes made to them have on the results. Rather, there is now the question of the extent to which the researcher can distance herself from everyday understandings of how search works, and instead employ the Web as a medium for place-based, language-based social research (i.e., using a search engine not as a consumer information appliance but as a research machine) [18].

Search as research practices have been already applied in studies of both the live and archived Web [7]. Some tools that support *search as research* are under development, however most state of the art studies are carried out manually, at least in part. The tools, Omnipedia [6] and Manypedia, for example, allow for the cross-cultural study of the 'same' article across language Wikipedias. Specifically, Omnipedia produces interactive diagrams that represent the similarities and differences existing among the language Wikipedias, thus showing salient information that is unique to each language edition. Manypedia [28] is a tool to compare the same Wikipedia page as it appears on 2 different language Wikipedias (both translated in the same language). Following Hecht and Gergle [10], the authors also calculate a level of similarity between two concepts presented in the pages, thus highlighting the "Linguistic Points of view (LPOV)" of different language Wikipedias.

Simple applications developed by the Digital Methods Initiative also allow for comparing images, tables of contents, editors of the 'same' article across language Wikipedias. Another social research tool, Contropedia, assesses the controversiality of an article [24].

The aim of this paper is to support this emerging area of research (and in particular cross-country or comparative media studies with web data) by providing an automatic and more comprehensive manner to carry out *Search as research* on the web.

3. THE SAR-WEB APPROACH

One important manner of approaching search results from a *search as research* perspective is visualizing engine outputs so as to reveal the contents in ways that distance the researcher from everyday search engine use, and open up the prospect of making findings as opposed to receiving information.

Inspired by the (largely manual) work of search engine comparison by the Digital Methods Initiative (DMI), we developed a web search tool (SaR-Web: Search as Research-Web) that allows the visualization of search results with a semantic added value in order to facilitate comparisons and further analysis.

The previous DMI study both relied upon and emphasized the engines' capacities to index, order and rank results, and, in an editorial approach, selected for the final output the top ten distinctive rights types, leaving them in the order that Google provided [18]. Those that were unique per country were highlighted; those that countries shared (and their respective rankings) also were marked as such.

In this paper, we adopt an automatic method using the Bing API to collect search results for the same query, "Rights", and put forward a semantic approach to compare results in four different languages, providing visual representations of search results, without the necessity to understand fluently all languages under investigation.

As in the original work we rely on the definition of the “local” as a result of the search engine local-domain functionality to carry out cross-country analysis. We use Bing as a search engine instead of Google because the Bing API is available for research (and Google’s, under its terms of service, is not). With respect to its research value, Bing can be compared to Google because it uses largely similar techniques to rank the search results (e.g., in their respective language and geographical spaces, users collectively co-author results through their previous clicking behaviour). This place-based and language-based crowd-sourcing is a crucial point that enables the very idea of search as societal research.

Even though Google has larger market shares (70 - 90 percent per country worldwide), meaning up to 4 billion people use Google, people using the Bing/Yahoo search engine number about 1/2 billion. This very large number is sufficient to investigate users’ behavior from Bing query logs for social research purposes, as has been done in previous reserach. In [21] query logs were studied in order to quantify the difficulties encountered by children of different ages in browsing and searching the Internet and to identify the topics they search for. In [22] query log files were analyzed to acquire information about users’ sessions and deduce how the language and search behavior of a user on a topic evolves over time.

An in-depth analysis of the differences between Google and Bing search results (including their geographical and linguistic specificities) would be important to document, and the same techniques we propose in this paper can be used to support such analysis, however much it is beyond the scope of this paper.

A further distinguishing feature of the SaR-Web approach concerns the query preparatory work. In order to prepare a normal search engine to be used for social research purposes, the researcher has to clean it of cookies, history and preferences, and uninstall toolbars which usually activate tracking mechanisms. So as to avoid that the user receives personalised results as opposed to generic ones, or at least ones overly affected by settings, SaR-Web in effect cleans the original search engine of personal settings, turning off localization and personalization. Apart from preparing it as a research machine, SaR-Web also does not track the user.

Finally, SaR-Web annotates the query results semantically, so that the researcher can compare results on the semantic level, and not only on the syntactic one.

4. EVOLUTION OF THE SAR-WEB PLATFORM

The SaR-Web platform is based on a previous release of the system called MWS-Web [1][5][19]. Originally, the Multimodal Web Search (MWS) platform has been designed to facilitate the study of the web as a corpus [2]. 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 [3]. In particular, MWS-Web provides an easy-to-use Web interface to configure the search engine for inclusion or exclusion of Web sites, localization and language settings, and all the query operators that are not necessarily known to users, including researchers. MWS-Web performs simultaneous searches for web, images and videos, and implements functionalities for saving and visualizing the search results. A key feature of MWS-Web is the capability to present search results not only as a result list as common search engines do, but also as pie charts that aggregate views and summaries.

Searches performed in MWS-Web also can be saved for subsequent analysis. Indeed, since the search results, retrieved from the Web, will change over time, it is important for the researcher to be able to save the current results.

Web objects (such as web pages, images, and videos) appearing in the result list can be selected and collected in a Storyboard, which is a visual narrative that accompanies the presentation of the results (similar to a Power Point presentation displaying the search results in a sequence). More details about the use of the storyboard in educational scenarios can be found in [19]. The use of the system in educational settings [12] has shown the potentialities of the system also as a tool to support research studies on the Web as a genre [4] as well as on the analysis of search results from a broader perspective.

Commercial search engines such as *Google* and *Yahoo* 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.

5. SEARCH AS RESEARCH WITH SAR-WEB

5.1 Overview

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. From a technical perspective 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. The workflow of the system is as follows. After a user search for a keyword, the returned results (URLs) are obtained from the Bing search and are sent to 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 sorted in descending order are retrieved in order to create the word cloud.

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

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

In detail, the main tasks performed by SaR-Web to achieve its functionality are as follows:

1. Localized search: the keyword 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 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 under investigation (Figure 1).

Figure 1 shows an example of a (semantic) word cloud for the search in the four languages related to the keyword, “Rights”. Note that the clouds do not contain words related to a specific language but concepts linked to the DBpedia ontology, thus allowing researchers to compare the results, overcoming the limit of cross language analysis based on text. When the corresponding concept is not present in DBPedia, we display a prefix at the beginning of the original term (see for example the word cloud including the German “de:Zugriffsrecht”, which has no corresponding entry in the English DBpedia).

By clicking on each concept the platform presents the list of query results connected to the specific concept shown in the cloud. In this way the researcher can see the original list of sources as they were ranked by the search engine, and check the full content of each document. In the current release of the system, colors of the words in the tag cloud do not play a semantic role.

SaR-Web allows the user to configure their searching by means of specific configuration parameters which allow users to perform their queries in the Web or News sphere, or to set the quality of the annotation extracted through the Dandelion API. Higher thresholds of the confidence value lead to fewer but usually more precise annotations.

Other parameters are related to:

- the option of specifying the number of results returned by the search engine to be included in the investigation;

- the option of including the full text of the web page and not only the title and the snippet;

- the option of excluding the results coming from popular (and overly recurring) web sites such as Wikipedia, YouTube and so on.

In the following sections we describe three examples in which the SaR-Web platform has been used to investigate the main DBpedia concepts (or keywords) related to the queries: “Rights”, “nuclear”, and “green tips”. When we use the term “concept” we refer to the entities described in DBpedia. From a *search as research* perspective, these terms are exemplary of different types of queries: “Rights” is an ambiguous or underspecified query, where one asks the search engine to return rankings of cultural practices or mores across the countries; “nuclear” and “green tips” are relatively unambiguous or specified queries where one asks the search engine to return common types of societal concern across the countries.

For each example, we provide the semantic word clouds in the four languages (English, Italian, French and German) as a visual representation of the results. In addition, we supply a table including the five most represented concepts (or keywords) for each language with their percentage, so as to make explicit the incidence of the keyword in the entire cloud (Table 2, 3, 4).

All the queries described in Section 5 have been executed across the Web sphere, by setting a confidence level of 0.6, and processing the first 20 results returned by the Bing search engine.

5.2 Query: “Rights”

The first example concerns the query “Rights”. Results in the four languages are illustrated as a cloud in Figure 1.

In Rogers` work [18] a list of rights *types* is reported (see table 1 left column).

In SaR-Web, the entity extraction procedure gives a more heterogeneous result of *keywords* (including some types) associated with the query “Rights”. From the semantic point of view the results show a good correspondence.

Rights types	**SaR-Web Concept**
English (USA): human rights	English: “Rights”
Italy: privacy rights	Italian: “Law”
Germany: children’s rights	German: “Zugriffsrecht” (access rights)
France: youth rights	French: “Human rights” “Children’s rights”

Table 1: Comparing rights

We notice a different distribution of relevance in the four word clouds: in English the concept “Rights” and in Italian the concept “Law” are very prominent compared to all other concepts in the cloud (they are respectively the 8% and 13% of the total in each language). In the German and in the French clouds the relevance of the words is more balanced. In German the most prominent terms do not have correspondence in DBPedia (“Benutzer”=user and “Zugriffsrecht”=access rights). In the French cloud, two concepts (“Children´s rights“ and “Human rights”) have the same relevance of 7% of the total).

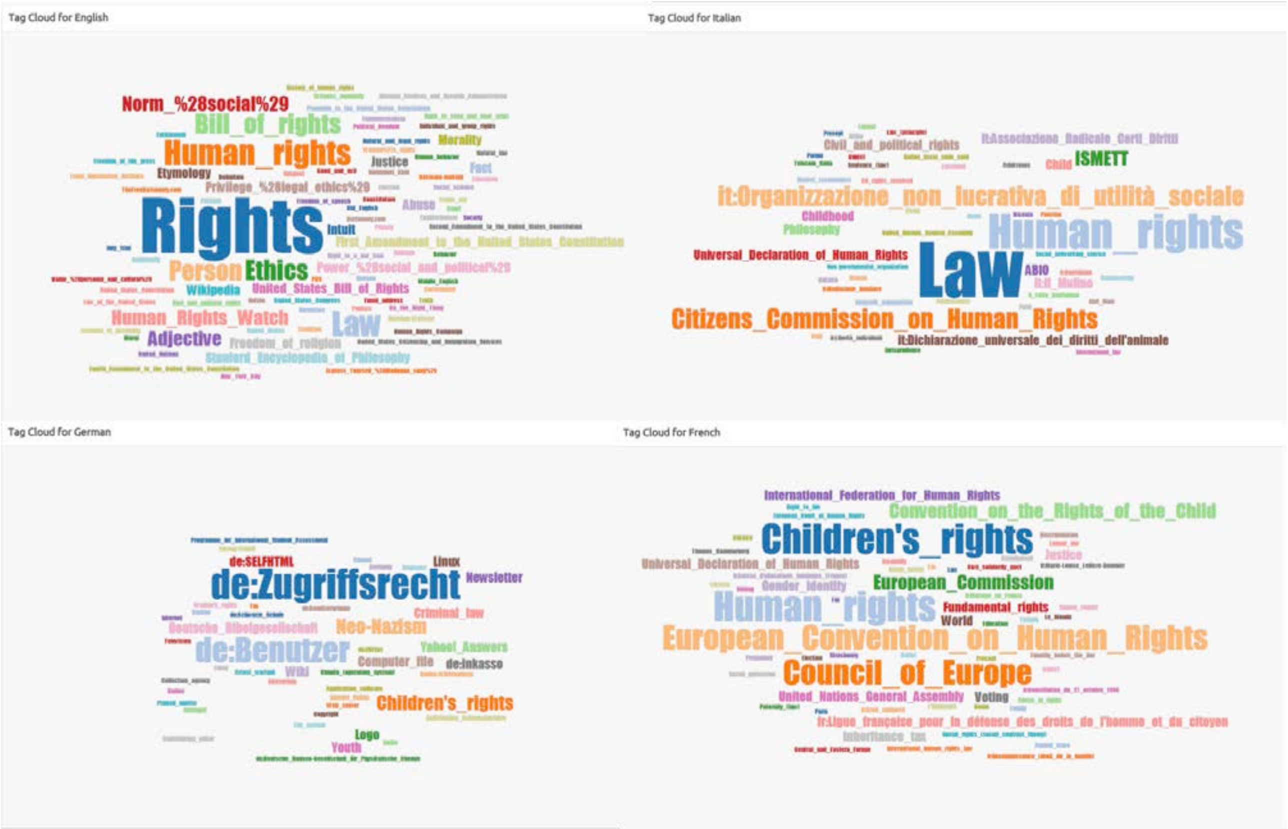


Figure 1: Concept Clouds for "Rights": clockwise from upper left ” in English, Italian, French and German

5.3 Query: “Nuclear”

In this example we enter the query ”Nuclear” (specified query) where we expect to obtain different results in terms of sentiments (for example, positive in France concerning nuclear energy and negative in Germany).

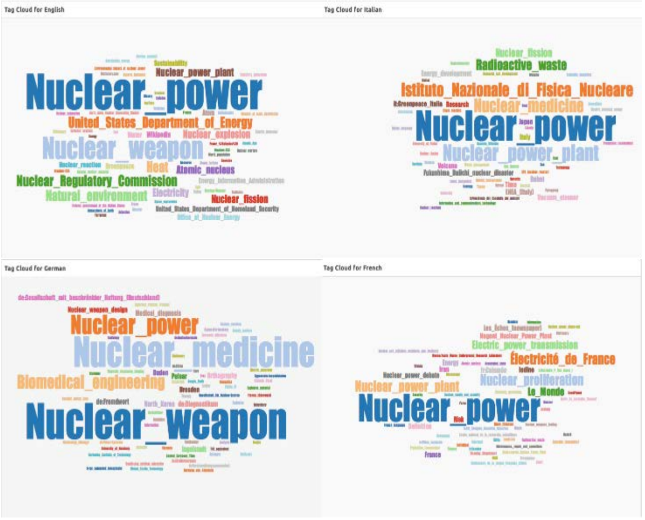


Figure 2: Concept Clouds for "Nuclear"[3]

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 (with a positive sentiment), while in the German cloud the most prominent concept, nuclear weapon, is associated to war (with a negative sentiment).

5.4 Query: “Green Tips”

The query “green tips” (Specified query) is likely to return culturally specified results per local domain, showing specified (but also shared) cultural practices (Figure 3).

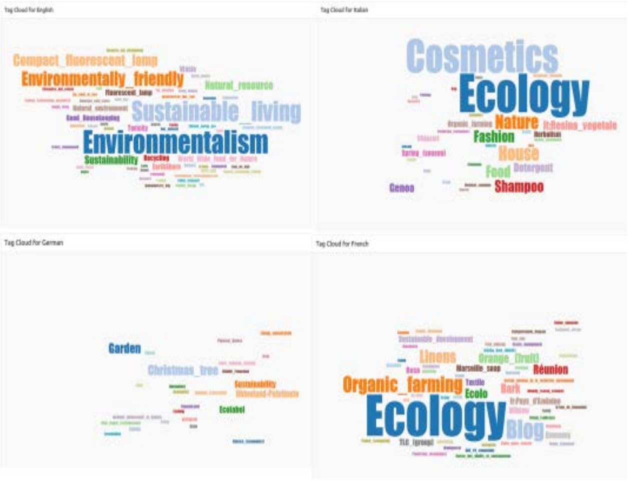


Figure 3: Concept Clouds for "Green tips"[4]

[3] Hi-res version of the figure is available at: http://pa.itd.cnr.it/websci16/nuclear.png

[4] Hi-res version of the figure is available at: http://pa.itd.cnr.it/websci16/greentips.png

Table 2: Query “rights, rechte, droits, diritti”

En		De		Fr		It	
Rights	8%	de:Zugriffsrecht	9%	Children's_rights	7%	Law	13%
Universal_Declaration_of_Human_Rights	6%	de:Benutzer	7%	Human_rights	7%	Human_rights	9%
Law	4%	Children's_rights	4%	European_Convention_on_Human_Rights	5%	Citizens_Commission_on_Human_Rights	5%
Human_rights	4%	Neo-Nazism	4%	Council_of_Europe	5%	ISMETT	5%
United_States_Bill_of_Rights	3%	Criminal_law	3%	Convention_on_the_Rights_of_the_Child	3%	Adolescence	4%

Table 3: Query “nuclear, nuklear, nucléaire, nucleare”

En		De		Fr		It	
Nuclear_power	15%	Nuclear_weapon	13%	Nuclear_power	11%	Nuclear_power	12%
Nuclear_weapon	7%	Nuclear_medicine	10%	Nuclear_proliferation	4%	Nuclear_power_plant	7%
United_States_Department_of_Energy	4%	Nuclear_power	6%	électricité_de_France	4%	Nuclear_medicine	5%
Natural_environment	3%	Biomedical_engineering	5%	Nuclear_power_plant	4%	Istituto_Nazionale_di_Fisica_Nucleare	5%
Nuclear_Regulatory_Commission	3%	Nuclear_weapon_design	2%	Electric_power_transmission	3%	Radiactive_waste	4%

Table 4: Query “green tips, grüne tipps, astuces écologique, trucchi ecologici”

En		De		Fr		It	
Environmentalism	9%	Garden	9%	Ecology	14%	Ecology	16%
Sustainable_living	8%	Christmas_tree	9%	Blog	6%	Cosmetics	13%
Environmentally_friendly	6%	Sustainability	6%	Organic_farming	5%	House	6%
Compact_fluorescent_lamp	4%	Rhineland-Palatinate	6%	Linens	4%	Nature	6%
Natural_resource	3%	Ecolabel	6%	Ecolo (Belgian political party based on green politics)	3%	Food	5%

In the Italian tag cloud we notice concepts related to beauty (e.g. Cosmetics, Fashion, Shampoo) which are less relevant in the other languages. Opening the associated original source in the local domain .it, we note that Cosmetics refers to a new trend of ecologically sound make up. The French cloud refers to such objects of ecological concern as linens, the Germans Christmas trees and the American a particular light bulb. In each case the absence of the cosmetics, linens, Christmas tree and light bulb in the other cultural space demonstrates the specificity of the particular discussions ongoing in each country, captured by the system.

6. DISCUSSION AND FURTHER EXPERIMENTS

Previous work on user behaviour for Web search showed that, most often, users tend to click on search results that appear on the first page of the search engine results page, rather than on the second or third page. In their study aimed at identifying individual skill related problems that users experience when navigating the Internet, Van Deursen and Van Dijk observed that 36% of users did not go beyond the first three search results, and 91% did not go further than the first page with search results [23].

The rank of a search result in the results page also influences the probability of the result being read and selected by the user. Through the Eye-tracking technique Granka et al. [9] investigated how users interact with the results page of a search engine, and found out that the time users spend on the first and second results of the list is on average the same, while it drops off sharply after the second result, and a further sharp drop occurs after link 10 (p. 479). Craswell and others report that the probability of a document being clicked depends on its rank in the results page, and the probability of observing a click decays with its rank [25].

Motivated by this previous work, we further investigated the results for the query “nuclear” in the German local domain (.de) which gave a different result in the previous experiment compared to the other languages (Figure 2). Using SaR-Web we set the parameters so as to consider only the first 10 search results in the local domain .de (according to the ranking in Bing) (Figure 4), and compared the results to those we obtained considering the first 20 results (Figure 2).

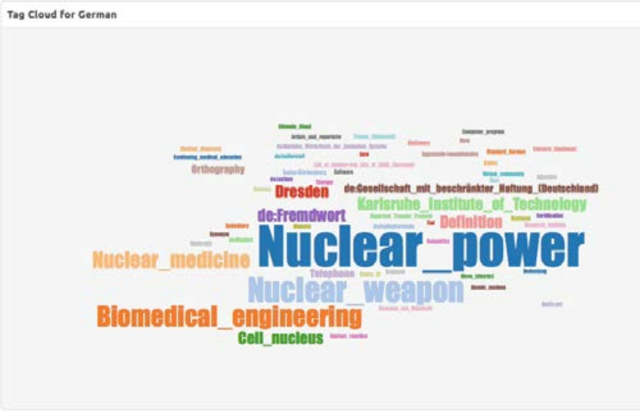


Figure 4: Query "Nuclear" for 10 results

Considering only the first 10 results corresponds to the behavior of a user who looks only at the first page of Bing results. In this case the most prominent concept associated to the keyword "nuclear" is Nuclear_power also in German (Figure 4), providing a somewhat more positive sentiment as opposed to the clearly negative one (Nuclear_weapon) which we obtained when we considered the first 20 search results (as we did in the preliminary comparison (Figure 2). This confirms that a crucial aspect to be taken into account for the analysis is the number of search results to be processed.

We also considered and processed 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 (Figure 6). The full text analysis provides more information which is useful to give an overview of the topic, but requires additional effort for the qualitative analysis of the results. For example it is not obvious why the concept cloud includes Dresden and the Karlsruhe Institute of Technology.

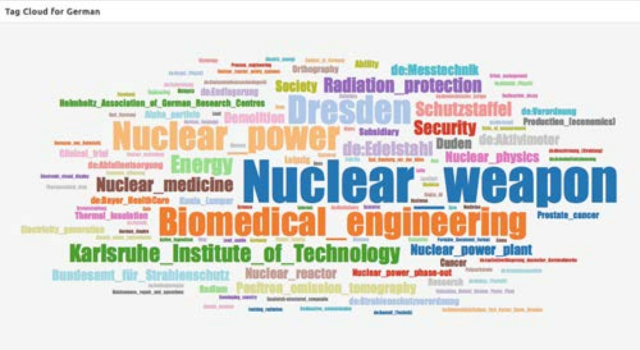
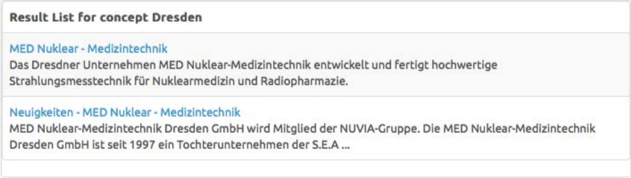


Figure 5: Nuclear Full-text - 10 results

SaR-Web also supports the researcher in this in-depth investigation by displaying the relevant results when the user clicks on a keyword in the word cloud. By clicking on the concept "Dresden" for example, the researcher realizes that both results mention the company MED Nuklear Medizintechnik based in Dresden.



Result List for concept Dresden

MED Nuklear - Medizintechnik
Das Dresdner Unternehmen MED Nuklear-Medizintechnik entwickelt und fertigt hochwertige Strahlungsmesstechnik für Nuklearmedizin und Radiopharmazie.

Neuigkeiten - MED Nuklear - Medizintechnik
MED Nuklear-Medizintechnik Dresden GmbH wird Mitglied der NUVIA-Gruppe. Die MED Nuklear-Medizintechnik Dresden GmbH ist seit 1997 ein Tochterunternehmen der S.E.A ...

Figure 6: Result List for the concept "Dresden"

Similarly Karlsruhe Institute of Technology appears because one result mentions a research program about nuclear safety performed by researchers at the university in Karlsruhe.

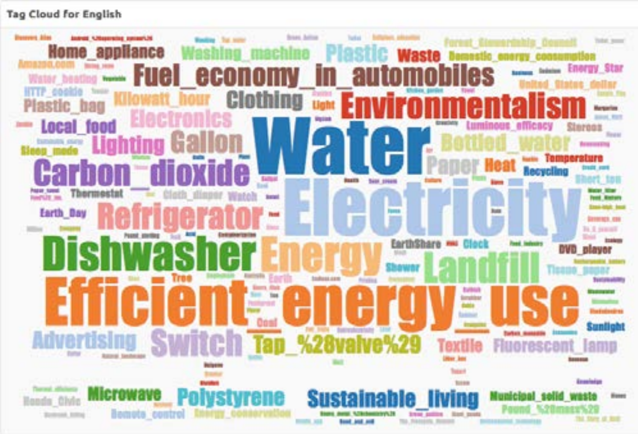


Figure 7: Query Green tips: full text with 10 results

Figure 7 provides the visualization from full text of the query "green tips". By comparing this picture with the visualization shown in figure 3 for English (which was generated only from snippets), we notice that both pictures give a very similar message (e.g. English people connect green tips with energy savings). Figure 7 gives more details, though, including useful examples suggesting that we can save energy using specific refrigerators or dishwashers, or by taking the fuel economy in automobiles into account. For these insights, we do not need to look at the specific webpages but can get these additional details already from the concept cloud itself.

7. CONCLUSIONS AND ONGOING WORK

Search as research provides methods for the social study of the Web and particularly the social meanings of engine results. *Search as research* techniques seek to transform a search engine from a consumer information appliance into a research machine. Such a research approach 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. Thus particular local results per query are boosted by users from the locations in question. *Search as research* seeks to take advantage of these 'research affordances' of engines through the use of specified and underspecified queries. How does each language or cultural space rank particular cultural practices (green tips) or societal concerns (nuclear energy or types of rights)? We have found that engines may reveal such practices and concerns through query designs that are able to tease them out.

Even though some automatic tools are available to support *search as research* practices, they are mainly limited to specific sub-activities or focus on specific knowledge repositories, such as Wikipedia. Consequently, these practices require manual work to prepare the search engine, specify local-domain settings for the country specificity of the languages, rank lists of results and refine the query.

In this paper, we have presented SaR-Web, a multimodal web search tool that provides an automatic way to carry out *search as research* on the Web. Inspired by the work of Richard Rogers and the Digital Methods Initiative, SaR-Web compares the results of same (translated) queries across search engine language domains, and visualizes search results with a semantic added value, thus facilitating cross-linguistic and cross-cultural comparisons of results.

SaR-Web features supporting *search as research* practices are still at a prototypical phase. In the next release of the system we plan to introduce additional functionalities to better facilitate the analysis of the results. First, the automatic search mechanisms implemented in SaR-Web can produce results which are out of scope of the original query, since there are no automatic filters to distinguish amongst website categories (e.g. online newspapers vs. academic sites). Second, results are sensitive to very recent events which could have temporary boosted some websites at the top of the engine results (which however is an interesting insight it itself as well), and pointing out event-related results will be an interesting extension in the future.

Another functionality under development concerns the possibility for the user to select a bilingual display mode, to compare the English version of the results of the query with the results in the original language. This visualization mode will improve the social analysis of the results (by highlighting ambiguities due to the representation of the concepts in the DBpedia ontology), and supports the use of the system in educational scenarios.

In order to improve the usability of the interface we plan to provide the following functionalities:

- ordered tag cloud (with a hierarchy of results);
- list results in an ordered table where the 4 countries are displayed side by side, with the possibility to edit and download the table;
- automatic translation of queries in various languages;
- filters to exclude specific items/concepts from the results;
- association of colours in the word cloud according to categories and concepts in the four different languages.

Moving from the manual inspection of engine results to automatic keyword extraction has been challenging, in that certain of the concepts in the database as well as those missing in the database would have been eliminated in an editorial approach. On the other hand, supporting search as research through automated means clearly makes it easier to analyze more queries and results, with the possibility of manual inspections to uncover additional insights. Thus the automated search as research work both should strive to continue to perfect the outputs, as well as being used as an intermediary step, prior to an editorial polishing. 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.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] Baldry, A.P. (2005) A multimodal approach to text studies in English. The role of MCA in multimodal concordancing and multimodal corpus linguistics. Campobasso: Palladino.

[2] Baldry, A.P. (2010). ‘A web-as-multimodal corpus approach to lexical studies based on intercultural and scalar principles’, in M. M. Jaén, F. S. Valverde and M. C. Pérez (eds.), *Exploring New Paths in Language Pedagogy. Lexis and corpus based language teaching.* London and Oakville: Equinox, pp. 173-190.

[3] Baldry, A.P. (2011). ‘Characterising transitions in identity in the Web: Multimodal approaches and methods’, in N. Vasta, A. Riem Natale, M. Bortoluzzi and D. Saidero (eds.) *Identities in Transition in the English-Speaking World,* dine: Forum Editrice Universitaria Udinese, pp. 17-38.

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

[5] Baldry, A.P., Gaggia, A. and Porta, M. (2011) “Multimodal Web Concordancing and Annotation. An overview of the MCAWEB System”, in Vasta, Nicoletta, Riem Natale, Antonella, Bortoluzzi Maria and Saidero Deborah (eds.) Identities in Transition in the English-Speaking World, Udine: Forum Editrice Universitaria Udinese, pp. 39-60.

[6] Bao P., Hecht B., Carton S., Quaderi M., Horn M., and Gergle D. (2012). Omnipedia: Bridging the Wikipedia Language Gap. *Proceedings of CHI ’12,* 1075-1084

[7] Ben-David, A., & Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. Alexandria, 25(1-2), 93-111.

[8] 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

[9] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '04). ACM, New York, NY, USA, 478-479. DOI=http://dx.doi.org/10.1145/1008992.1009079

[10] Hecht, B. and Gergle, D. (2010). The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. *Proceedings of CHI ’10,* 291-300

[11] Chen J. and Bao Y. (2009). Information access across languages on the Web: from search engines to digital libraries. Proceedings of the American Society for Information Science and Technology, 46(1):1–14.

[12] Kantz, D. (2012) ‘Medical CLIL (Part III): How the mind works’, in Cambria, Mariavita and Arizzi, Cristina and Coccetta, Francesca (eds). Web Genres and Web Tools. Como: Ibis, pp. 379-390.

[13] 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.

[14] Notess, G. R. (2008). Multilingual Searching: Search Engine Language Tools. Online, May/June 2008, 32(3), p40-42. retrieved September 1, 2008 at: http://www.infotoday.com/ONLINE/may08/Notess.shtml

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

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

[17] Ali Hosseinzadeh Vahid, Piyush Arora, Qun Liu, Gareth J. F. Jones (2015). A Comparative Study of Online Translation Services for Cross Language Information Retrieval. WWW 2015 Companion, May 18–22, 2015, Florence, Italy. ACM 978-1-4503-3473-0/15/05. http://dx.doi.org/10.1145/2740908.2743008

[18] 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.

[19] 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

[20] Zhang, J. & Lin, S. (2007). Multiple language supports in search engines. Online Information Review, 31(4), 516-532. Retrieved August 18, 2008, from http://www.emeraldinsight.com/Insight/ViewContentServlet;jsessionid=26559230F1B5E51B9C81A07DB8D54796?Filename=Published/EmeraldFullTextArticle/Articles/2640310408.html.

[21] Duarte Torres, S., Weber, I., and Hiemstra, D. (2014). Analysis of search and browsing behavior of young users on the web. ACM Trans. Web 8, 2, Article 7 (March 2014), 54 pages. DOI: http://dx.doi.org/10.1145/2555595

[22] Eickhoff C., Teevan J., White R., and Dumais S. (2014). Lessons from the journey: a query log analysis of within-session learning. In Proceedings of the 7th ACM international conference on Web search and data mining (WSDM '14). ACM, New York, NY, USA, 223-232. DOI=http://dx.doi.org/10.1145/2556195.2556217

[23] 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.

[24] E. Borra, E. Weltevrede, P. Ciuccarelli, A. Kaltenbrunner, D. Laniado, G. Magni, M. Mauri, R. Rogers & T. Venturini. " Contropedia – the analysis and visualization of controversies in Wikipedia articles." In Proceedings of the 10th International Symposium on Open Collaboration (OpenSym 2014). New York: ACM, 2014, 2014:34.

[25] Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008, February). An experimental comparison of click position-bias models. In Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 87-94). ACM.

[26] Chen, J. (2006). A lexical knowledge base approach for English-Chinese cross-language information retrieval. Journal of the American Society for Information Science and Technology, 57(2), 233-243.

[27] Oard, D. W., & Diekema, A. R. (1999). Cross-language information retrieval. In M. Williams (Ed.), Annual Review of Information Science and Technology, 33 (pp. 223-256).

[28] Massa P. and Scrinzi F. 2012. Manypedia: comparing language points of view of Wikipedia communities. In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym '12). ACM, New York, NY, USA, , Article 21 , 9 pages. DOI=http://dx.doi.org/10.1145/2462932.2462960