Inf Retrieval (2009) 12:324–351
DOI 10.1007/s10791-008-9082-8

Mixed monolingual homepage finding in 34 languages: the role of language script and search domain

Roi Blanco · Christina Lioma

Received: 19 May 2008 / Accepted: 27 November 2008 / Published online: 20 December 2008
© Springer Science+Business Media, LLC 2008

Abstract The information that is available or sought on the World Wide Web (Web) is increasingly multilingual. Information Retrieval systems, such as the freely available search engines on the Web, need to provide fair and equal access to this information, regardless of the language in which a query is written or where the query is posted from. In this work, we ask two questions: How do existing state of the art search engines deal with languages written in different alphabets (scripts)? Do local language-based search domains actually facilitate access to information? We conduct a thorough study on the effect of multilingual queries for homepage finding, where the aim of the retrieval system is to return only one document, namely the homepage described in the query. We evaluate the effect of multilingual queries in retrieval performance with regard to (i) the alphabet in which the queries are written (e.g., Latin, Russian, Arabic), and (ii) the language domain where the queries are posted (e.g., google.com, google.fr). We query four major freely available search engines with 764 queries in 34 different languages, and look for the correct homepage in the top retrieved results. In order to have fair multilingual experimental settings, we use an ontology that is comparable across languages and also representative of realistic Web searches: football premier leagues in different countries; the official team name represents our query, and the official team homepage represents the document to be retrieved. A series of thorough experiments involving over 10,000 runs, with queries both in their correct and in Latin characters, and also using both global-domain and local-domain searches, reveal that queries issued in the correct script of a language are more likely to be found and ranked in the top 3, while queries in non-Latin script languages which are however issued in Latin script are less likely to be found; also, queries issued to the correct local domain of a search engine, e.g., French queries to yahoo.fr, are likely to have better retrieval performance than queries issued to the global

R. Blanco (✉)
Computer Science Department, University of A Coruna, La Coruña, Spain
e-mail: rblanco@udc.es

C. Lioma
Computer Science Department, Katholieke Universiteit Leuven, Heverlee, Belgium
e-mail: christina.lioma@cs.kuleuven.be

domain of a search engine. To our knowledge, this is the first Web retrieval study that uses such a wide range of languages.

Keywords Multilingual information retrieval · Web information retrieval · Search engines and evaluation

1 Introduction

In 2005, World Wide Web inventor Berners-Lee stated that universality is the key property of the Web: One must be able to access the Web whatever the hardware device, software platform, and network used, whether one is in a "developed" or "developing" country, and that information must be supported in any language, culture, and field without discrimination (Berners-Lee 2005). An example of technology that enables access to the universal Web is Information Retrieval (IR) systems. IR systems aim to retrieve information, which is relevant to a user need, from a given repository of information, such as a document collection (Van Rijsbergen 1979). A common application of IR systems is Web search engines, in which a short keyword query is used to generate a ranked list from a previously indexed part of the Web.

Since its conception in 1991, the Web has come a long way. Its size, popularity, spread and integration into society far exceed even the wildest expectations, with 170 million Websites and more than 20 billion indexed pages as of March 2008.[1] Whereas initially the Web was the domain of a few select mainly English-speaking computer users, and "terra incognita" to both industry and academia, nowadays research and business are turning to it, while estimates refer to hundreds of billions of Web users worldwide, the majority of whom are non-native English speakers. The amount of non-English speaking Web users continues to grow rapidly, and probably faster than it does for English speakers (Lazarinis et al. 2007), as increasingly more non-English speaking parts of the world go "online"; e.g., Cubans just recently got access to computers, it is a matter of time before they can access the Web, while statistical projections estimate that Chinese Web users are about to outnumber English speaking Web users (Fallows 2007). Internet usage is in fact reported to be 29.5% English, 70.5% non-English (Efthimiadis et al. 2007). This means that information available or sought on the Web is multilingual, and that technology providing access to this information needs to deal with this. To address this need, research into multilingual information access has been rigorously pursued, e.g., by the Cross-Language Evaluation Forum (CLEF, http://www.clef-campaign.org/) and NTCIR Asian Language Retrieval and Question-Answering Workshop (http://research.nii.ac.jp/ntcir/), as well as SIGIR and ACL workshops (Lazarinis et al. 2007; Workshops on Multi-source, Multilingual Information Extraction and Summarization (MMIES, http://doremi.cs.helsinki.fi/mmies2/)). The industry has also responded, for instance by embracing new standards in computer character encoding, namely the new standard Unicode 5.1, which contains over 100,000 characters, and provides significant additions and improvements that extend text processing for software worldwide (http://googleblog.blogspot.com/2008/05/moving-to-unicode-51.html). These initiatives show the effort of academia and industry to deal with multilingual digital information.

[1] Infoscale 2008 keynote by R. Baeza-Yates: http://www.infoscale.org/keynote.shtml.

The increase of Web multilinguality has raised the question of how well existing technology can deal with non-English information, since most search engines were originally engineered for English (Lazarinis et al. 2007). For instance, it is commonly held that encoding, capitalisation, diacritics, compounding, complicated morphology and other language-specific issues can limit the performance of search engines in retrieving multilingual information (Gey et al. 2006). In addition, it is usually acknowledged that international search engines are less effective with non-English than with English queries (Macdonald et al. 2007).

In this work, we study how existing search engines respond to the challenge posed by multilingual queries, by asking these research questions:

1. Can different language alphabets (*script*) affect retrieval performance, and if so to what extent? This question is motivated by existing literature on script issues in multilingual information access (Ahmed et al. 2007; Chung 2008; Efthimiadis et al. 2007; Hasan et al. 2000; Lazarinis 2007; Tzekou et al. 2007), and also by the lack of standardisation in transcribing & transliterating non-Latin script languages.
2. Does limiting the search domain from global (e.g., google.com) to local (e.g., google.fr) improve retrieval performance? This question is motivated by the propagation of local domains in most major search engines (e.g., google.it., msn.fr, yahoo.es), as well as recent IR evaluation initiatives into European domain-based retrieval through WebCLEF (http://ilps.science.uva.nl/webclef/).

We address these research questions in the context of homepage finding, a standard Web IR task, in a mixed monolingual setting, a standard WebCLEF task. Given a corpus, e.g., the Web, and a set of queries with associated a unique homepage each, the retrieval task is to return that homepage and rank it as high (i.e., relevant) as possible. This task corresponds to the scenario of a Web user who wants to visit the Webpage of an entity he knows, but who does not know where to find that page (Craswell et al. 2001). We select the realistic scenario of Web users who want to visit the Webpage of a football team; we consider as queries the names of teams competing in the premier leagues of 71 different nations according to the Fédération Internationale de Football Association (FIFA, http://www.fifa.com/). This provides us with an ontology, where the entities consist of different teams; these entities are structured according to nation and also with respect to their ranking in the national premier league. This ontology is comparable across a wide range of languages, and also allows us to generate queries in different languages in the exact same way (using the team name). Thorough experiments involving over 10,000 runs in 34 different languages using four major freely available Web search engines reveal that queries issued in the correct language script have better early precision (if the homepage is retrieved, it will be ranked in the top 3), while queries of a non-Latin script language issued in the Latin script are less likely to be found. In addition, querying local domains has better overall retrieval performance than querying global domains, for most languages.

The main contributions of this work are:

– Two known issues in multilingual IR (language script (Ahmed et al. 2007; Chung 2008; Efthimiadis et al. 2007; Hasan et al. 2000; Lazarinis 2007; Tzekou et al. 2007) and search domain (Balog et al. 2007)) are addressed using realistic settings, e.g., the Web as a collection, and major search engines as IR systems. To our knowledge, the problem of language script is not addressed in any organised or principled way so far.
– With 71 domains, 34 languages and 764 queries, and variations of these with respect to script and domain, this is a thorough investigation, the scale of which is unprecedented

to our knowledge in the field of multilingual Web retrieval: WebCLEF uses 24 languages and 27 domains, only Europe-wide (Sigurbjörnsson et al. 2006); other CLEF tasks (e.g., ad-hoc search) use even fewer languages, and not for Web retrieval.
- The use of FIFA premier league data provides an ontology suited to the mixed monolingual Web IR task, which represents a realistic search scenario, and also a novel way of mining comparable data from freely available sources. We have made freely available the complete query—correct answer dataset.[2]
- This is one of the few Web IR studies using Bosnian, Croatian, Estonian, Hebrew, Slovene, Slovak, Vietnamese, and the first IR study using Albanian, Armenian, Azerki, and Uzbeki, to our knowledge.

The rest of this paper is organised as follows. Section 2 motivates this work, and Sect. 3 presents the state of the art in multilingual Web retrieval. Section 4 introduces our step-by-step methodology for addressing our research questions. Section 5 discusses the settings and outcome of our investigation. Section 6 concludes this paper and states future research directions.

2 Motivation

Different languages are written in different alphabets or script, e.g., English and Italian use Latin script, Arabic[3] uses Arabic script, and so on. The predominance of English on the Web has also resulted in a predominance of the Latin script on the Web, meaning that languages that do not use Latin script are often written in Latin script, for instance by non-speakers of that language, or out of convenience. For example 'Greeklish' is the term referred to Greek written in Latin script (Tzekou et al. 2007). This process is very similar to the *transliteration* and *transcription* processes, which refer to writing words in one language using the alphabet of another language, not translating them. Broadly speaking, transliteration is the spelling of words from one language with characters from the alphabet of another, usually in a character-by-character replacement process (Garfield 1975); whereas transcription is the representation of the *sound* of words in a language using any set of symbols, i.e., the alphabet of another language or the International Phonetic Alphabet (IPA) (Garfield 1975). There can be various different transliteration and transcription outputs, for instance the Russian name `Khrushchev` (here transliterated by an English speaker) can also be transliterated as `Chrustschev` by German speakers; the same name can also be transcribed as `Khrooshtchoff` by English speakers, `Chruhszhtchow` by German speakers, `Jruchev` by Spanish speakers, `Chroesjtjov` by Dutch speakers, or `Crustsciof` by Italian speakers. In official transliteration and transcription by organised bodies, e.g., visa issuing offices, there are standards to be adhered to, but on the Web there are no standards on how a word will be written in a script other than its own. Very often, a mixture of transcription and transliteration practices are combined, also assisted by the use of numbers and punctuation. For example, the Greek letter θ is often written as the number '8'.

Writing non-Latin languages in Latin script is an issue because the same language entities are represented under different forms, i.e., no new words are added to the language, only different ways of writing the same words. For IR systems this is both an indexing and

[2] Freely available upon request from the authors.

[3] In this work, by 'Arabic' we refer collectively to all variants that share a more or less common writing form, e.g., Egyptian, Algerian, and so on.

a matching problem: For example, a Russian term can be written in Russian letters, and also in Latin letters in multiple ways; should all these term variants be indexed as one entry or as separate entries? Should these terms be normalised in some way, e.g., stemmed? Should a query containing the term in Russian letters be matched to a relevant document containing the term in Latin letters? Should the term written in Russian letters be weighted identically as the same term written in Latin letters? Should search engines provide advanced search options for the automatic transcription of queries? The problem is even more accentuated for languages with ideographic characters, like Chinese[4] and Arabic for instance, where there is no one-to-one correspondence with Latin letters. For example, the standard Hebrew[5] orthography leaves most vowels unspecified: It does not explicate 'a' and 'e', does not distinguish between 'o' ad 'u', and leaves many of the 'i' vowels unspecified. Furthermore, the single letter 'w' is used for all 'o', 'u' and 'v' (Daya et al. 2008). In such languages, tokenisation is required, which can be costly and can also introduce error to the whole of the retrieval process (Hasan et al. 2000).

To our knowledge, this problem of writing non-Latin languages in Latin script is not addressed in any organised or principled way so far. This is not surprising, given the number of languages that become increasingly present on the Web, and the free and very creative way in which users express themselves. By 'free and creative' we mean that the problem of correct language script does not occur in isolation on its own, but often combined with other known problems of 'Web linguistics', such as using abbreviations or slang, reversing syllables or replacing letters in words, or coining new terms and concepts, e.g., `bonk`, `moblog`, `netspeak`[6] (see Crystal 2006; Mishne 2007, for an overview). This is a hard task for a machine to resolve.

Our second research question, namely the use of local language-based search domains, is motivated by the observation that increasingly more search engines make available local language-based search domains. For instance, today Google offers over 65 different local language domains, in addition to google.com. This proliferation of language-based search portals by major search engines has caught the attention of the scientific world: In the May 2008 issue of the ACM Communications, language-based search engine domains are considered as having great potential for 'opening information treasures' (Chung 2008). We attempt to evaluate and validate this potential, by asking: What do language-based search domains offer on top of global search domains, if any? Should users use them? Furthermore, is domain-based search equally effective for all languages?

3 Related studies

The mixed monolingual homepage finding search task studied in this work is similar to the mixed monolingual Web retrieval track of WebCLEF (Balog et al. 2007). The WebCLEF retrieval task is based on a stream of known-item topics in a range of languages, which have to be retrieved from the EUROGOV collection, a collection of spidered Web sites of European governments (Sigurbjörnsson et al. 2006). EUROGOV contains webpages from

[4] In this work, by 'Chinese' we refer collectively to all variants that share a more or less common writing form, e.g., Cantonese, Mandarin, and so on.

[5] In *undotted* or *unvocalised* script only, which is standard (Daya et al. 2008).

[6] Bonk = a mistyped character that changes the meaning of a word usually into something naughty. Moblog = a blog of posts mainly sent by mobile phone. Netspeak = the language of the Web (Crystal 2006).

27 domains, covering over 20 languages. The conclusion of the WebCLEF evaluation so far has been that participating IR systems are very effective, retrieving on average the target page in the top ranks. Our study follows the WebCLEF evaluation to an extent: The task is mixed monolingual homepage finding and evaluation is based on the rank of the retrieved results. However, there are several differences from WebCLEF: First, the search space queried in WebCLEF is European and restricted to the .gov domain, whereas we query the whole Web. Second, WebCLEF queries are available in around 20 languages, whereas our queries cover 34 languages. Third, the nature of the queries differs, mainly governmental for WebCLEF and football-oriented for our study. Even within these domains however, WebCLEF queries cover significantly more topics than our queries, which consist solely of football team names, hence our study is certainly more restricted than WebCLEF in this respect. Fourth, WebCLEF queries are more verbose than our queries, for instance they include a description field that elaborates on the information need, as well as several metadata describing for example the language of the query or the domain of the target page. On the contrary, our queries are very short, consisting of the team's name only.

More generally, the task of homepage finding is a typical search task on the Web (Broder 2002), and as such it has been studied extensively, for instance as part of the recent Text REtrieval Conference (TREC) (Voorhees et al. 2005) Web track (Craswell et al. 2001) and Terabyte track (Buttcher et al. 2006; Buttcher et al. 2005), for the English language. These evaluations used controlled datasets, i.e., collections crawled from the general Web (WT2G, WT10G) and the .gov domain (.GOV2) respectively, and concluded that IR systems were able to return the correct answers and rank them high in most cases; in 2005 and 2006 the rate of not found Web pages was 17.1% and 12.7% respectively, quite low in both cases, meaning that most pages were found; regarding the rank of the found pages, measured in mean reciprocal rank (explained in Sect. 5.1), best scores in 2005 and 2006 were 0.463 and 0.512 respectively. In Sect. 5.2, we show how these scores, achieved by experimental systems in controlled monolingual experimental settings, compare to the scores we report in this paper of major search engines operating live on the multilingual Web.

Regarding the retrieval techniques used in multilingual retrieval, these consist mostly of extending mainly English IR systems into performing retrieval in or between other languages: one common technique is to use Web-based features (for example document structure) in order to facilitate retrieval in a multilingual setting (Adriani et al. 2006; Figuerola et al. 2006; Heuwing et al. 2007; Macdonald et al. 2006; Martínez-González et al. 2006; Tomlinson et al. 2006); another technique is to expand queries with assumed relevant terms (also known as pseudo relevance feedback) (Rodríguez et al. 2007; Santiago et al. 2007; Tomlinson et al. 2007; Wijaya et al. 2007) or more generally to reformulate queries in order to render them more informative (Balog et al. 2007); language-specific stemming is also common when retrieving documents in different languages (Adriani et al. 2007; Macdonald et al. 2006; Orengo et al. 2007; Tomlinson et al. 2006); another language-specific technique sometimes used in the area is decompounding, which consists of splitting compound words into their respective counterparts in order to facilitate the indexing and matching of terms (Savoy et al. 2007); an alternative to stemming and decompounding in a multilingual environment is the use of character n-grams to represent the terms in the index (Jensen et al. 2006; McNamee et al. 2007; Vilares et al. 2007); further techniques used with retrieval in different languages include adjustments to the retrieval ranking functions by reducing the term space used for matching documents to queries (López et al. 2007), or by applying penalising constraints to ranking (Pinto et al.

2007), as well as the combination of different indices (Balog et al. 2007); normalising diacritics and accents is also a known issue in the area (Kamps et al. 2006); finally, encoding issues, one of the biggest problems with non-English retrieval, have been dealt with either by adapting the retrieval system to process specific encodings, such as UTF-8 for example (Macdonald et al. 2006), or by transliterating characters into encodings that the system can process (Kamps et al. 2006).

In terms of multilingual IR, our work addresses mixed monolingual retrieval, not crosslingual retrieval. Broadly speaking, in crosslingual retrieval the query and retrieved documents are in different languages, hence the main issue is how to translate either or both, whereas in mixed monolingual retrieval the query and retrieved documents are in the same language, hence the main issue is how to identify the language and retrieve efficiently in it. The area of crosslingual IR is studied extensively, for instance within CLEF and NTCIR. The area of mixed monolingual IR is studied extensively in WebCLEF. Finally, regarding non-English monolingual IR in particular, increasingly more initiatives appear, for instance the recent Workshop on Improving Non-English Web Searching (iNEWS), which also concluded that IR systems need to consider more language-specific aspects (Lazarinis et al. 2007).

4 Methodology

Our experimental methodology for addressing our two research questions has four main steps:

1. *Choose corpus* We choose the Web.
2. *Identify query pairs* Identify a set of <query, homepage> pairs (for example <Real Madrid, http://www.realmadrid.com/>). Each pair describes the user's query (team name) and underlying need (team homepage). The query is the official team name as used by FIFA. The correct answer is the uniform resource locator (url) of the homepage.
3. *Query the corpus* For each research question, run the queries over the corpus using a freely available search engine.
4. *Measure effectiveness* Apply some retrieval effectiveness measure. In case of multiple equivalent correct answers, measure according to the top ranked one.

For Step 2, considering the team name as the query is the standard form of most navigational queries in homepage finding (Craswell et al. 2001). If there are more than one urls corresponding to the team homepage, e.g., mirrorred urls, any of them is the correct answer.

For Step 3, we query the corpus to address our research questions as follows:

– *Question 1: effect of language script on retrieval*

   1. *Correct script* we use queries in the correct language script, e.g., English queries written in Latin letters, Armenian queries written in Armenian letters, and so on;
   2. *Latin script* we use queries in Latin letters only, and exactly as the team names are referred to by FIFA. We choose the FIFA names to avoid making any transcription or transliteration decisions, which might bias the experiments, and also which might be questionable given that there are no transliteration or transcription standards.

– *Question 2: effect of search domain on retrieval*

1. *Global domain* we query the global domain of the search engine, for instance, google.com;
2. *Correct local domain* we query the local domain of the search engine corresponding to the language of the query, e.g., google.fr for French queries, yahoo.de for German queries, and so on. We skip languages for which no search engine has a local domain.
3. *Incorrect local domain* we query the local domain of the search engine corresponding to a different language than the language of the query. For this experiment, we wish to test the hardest retrieval scenario where the language of the query is very different to the language of the local domain, e.g., Russian queries in the Japanese local domain. We select the query language—domain language pairs as follows:

   (a) we arrange languages into groups, according to the language family they belong to; this reduces our 34 languages into 5 main language groups (see Sect. 5.1); we do this because languages of the same family often share many common features, especially words, so querying within languages of the same family (e.g., Italian–Spanish) is an easier task than querying between languages of different families (e.g., Italian–Arabic).
   (b) from each language group, we select one language, which is 'difficult' and also for which search engines have a local domain. We define a language as difficult if it has complicated morphology and uses non-Latin script, if possible. E.g., we select Chinese from the Sino-Tibetan language group, and Russian from the Indo-European language group;
   (c) we use the local domains of each of the 5 selected languages to run our queries, but excluding runs within the same language group. E.g., we run Chinese queries to the Russian domain, but not to the Chinese domain.

So, to address Research Question 2, we submit queries first to a global domain, then to their correct local domain, and then to four different incorrect local domains.

In order to analyse all combinations of the encoding and domain issues examined, we conduct experiments for Question 1 using separately each of the variations in Question 2. We expect the correct domains to perform better than the incorrect domains.

Note that our methodology is not without limitations: It does not address homepage finding scenarios where the search is biased, by factors such as the requirement for special software or keyboard in order to issue queries in non-Latin text, or where there may exist multiple domains for a homepage and in different languages, which may help the search engine. Moreover, with respect to language bias, our use of the FIFA domain can also restrict this study, because considering FIFA team names as the official team names, or adopting FIFA's rendering of non-Latin script are choices that may be questioned, especially from a language perspective.

5 Evaluation

5.1 Evaluation settings

We focus on the task of homepage finding, in the context of sports and specifically football. Generally, searching for football information is a very realistic search scenario, as

witnessed by the recent Google functionality of applying named entity recognition automatically to display information of upcoming matches or scores of ongoing matches (http://googlesystem.blogspot.com/2008/05/google-onebox-for-premier-league.html). Our retrieval collection is the Web, and our retrieval system is four different major commercial search engines: Ask (http://www.ask.com/), Google (http://www.google.com/), Microsoft's Live Search (former MSN Search, http://www.msn.com/), Yahoo! (http://www.yahoo.com/). We do not have an insight into the indexing and retrieval strategies of these engines, so we treat them as a black box: We submit queries to them and analyse the returned results. Our queries are names of football teams which compete in their national premier league in 2008 according to FIFA, and which also have a homepage on the Web. Teams without a homepage were excluded.

In total we collected 764 unique queries (teams) from 80 different national football leagues. We grouped queries according to language, e.g., we grouped English and Irish queries under English, and Chilean and Spanish queries under Spanish. This resulted in 71 different domains and 34 different languages. Query length was short (1.85 words on average, which is realistic of this task) and consistent between languages (average query length per language ranged between 1.1 and 2.7, see Table 1.[7] Table 1 also shows the percentage of native speakers of a language who use the Web. We see that there is no bias in our experiments towards having more queries for languages with more Web users, because the number of queries used depended only on the number of premier league football teams with Webpages; for instance, we used <20 queries for Danish, Dutch, and Swedish (for which >50% of native speakers are also Web users), and we used >30 queries for Spanish, Portuguese, and Arabic (where <20% of native speakers are Web users). The only bias of our dataset is towards languages with more premier league football teams with a presence on the Web. In this respect, our query dataset is domain-restricted and hence results between languages are not directly comparable outside this domain.

For the incorrect local search domain experiment described in Sect. 4, we grouped queries into five main language families (Voegelin et al. 1977), and selected a 'difficult' language from each family, shown in bold in Table 2. Overall, we submitted more than 10,000 runs, with more than 4,000 runs to the global or the correct local domain, and the rest to the local incorrect domain experiment. The scale of the experiments is shown in Fig. 1. The difference in the number of queries submitted to the global and local domains by different search engines is due to the fact that not all search engines had local domains for all languages; the local domains reported here are a snapshot of the period April-May 2008.

Web search engines may redirect user queries initially posted to their global engine (e.g., ask.com) to their local country pages (e.g., ask.es) on the basis of the internet protocol (ip) address of the incoming hypertext transfer protocol (http) request. This restriction can be easily surpassed for Ask, Google, and Yahoo!, but not for Microsoft's Live Search. For this reason, we selected the USA Live portal as Live's global (.com) page, which is likely to have indexed the bigger amount of webpages. Note that the redirection is not applied if the query is sent to a local domain page.

In our evaluation, we tried to detect all duplicate, mirrored, or redirected correct homepages by doing a string search on the path of the url address. E.g., we considered correct both: "http://www.realmadrid.com", "http://www.realmadrid.com/index.php". In addition, we tried to manually detect alternative spellings in the url paths, for instance

[7] Source of native speaker statistics presented in Table 1: http://global-reach.biz/globstats/index.php3 and http://en.wikipedia.org/.

Table 1 Number of queries per language and average number of words per query

Language (Web users)	Nation(s)	Queries	Avg. words
English (84.3%)	Australia, Belize, Canada, Cayman Islands, England, Ireland, Northern Ireland, USA, Scotland, Wales, Zimbabwe	114	2.0
Spanish (18.3%)	Argentina, Bolivia, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Mexico, Spain, Venezuela	102	1.9
Portuguese (10.4%)	Brazil, Portugal	56	1.8
French (24.8%)	Belgium, Cameroon, Cote d'Ivoire, France, Luxembourg	51	1.9
Arabic (2.5%)	Algeria, Bahrain, Egypt, Jordan, Lebanon, Sudan, Syria	35	1.7
German (52.4%)	Austria, Germany	28	2.7
Russian (11.1%)	Belarus, Russia	27	2.1
Greek (22.5%)	Cyprus, Greece	26	1.6
Italian (39.2%)	Italy	20	1.1
Dutch (54%)	Aruba, Netherlands	19	1.3
Japanese (54.5%)	Japan	18	2.3
Turkish (7.8%)	Turkey	18	1.3
Romanian (9.1%)	Romania	17	2.1
Hungarian (12.1%)	Hungary	16	1.4
Iranian (no stats)	Iran	16	1.2
Polish (18.3%)	Poland	16	2.0
Swedish (87.8%)	Sweden	16	1.9
Czech (35%)	Czech Republic	15	2.2
Bulgarian (no stats)	Bulgaria	14	2.1
Finnish (45.9%)	Finland	14	2.5
Slovak (21.4%)	Slovakia	12	2.2
Croatian (16.1%)	Croatia	11	1.7
Danish (62.5%)	Denmark	11	1.8
Chinese (8.0%)	China	10	2.2
Estonian (no stats)	Estonia	10	2.3
Hebrew (38.3%)	Israel	10	2.0
Albanian (no stats)	Albania	9	1.4
Bosnian (no stats)	Bosnia	9	2.0
Slovenian (no stats)	Slovenia	9	1.9
Vietnamese (3.1%)	Vietnam	8	1.7
Lithuanian (no stats)	Lithuania	7	2.0
Azerki (no stats)	Azerbaijan	6	2.2
Armenian (no stats)	Armenia	4	2.0
Uzbeki (no stats)	Uzbekistan	4	2.0
All		764	1.85

In brackets % of native speakers of a language who use the Web

"http://www.olympiakos.gr" and "http://www.olympiacos.gr", when possible. Our approach might have left correct duplicate homepages undetected. A more principled approach than our string search and manual inspection, to be considered in the future, would be to identify duplicate homepages using the National Institute of Standards and

Table 2 Language families

Language family	Languages
1. Asiatic	Arabic, **Hebrew**, Vietnamese
2. Indo-European	Albanian, Armenian, Bosnian, Bulgarian, Croatian, Czech, Danish Dutch, English, French, German, Greek, Iranian, Italian Lithuanian, Polish, Portuguese, Romanian, **Russian**, Slovak, Slovenian, Spanish, Swedish
3. Japanese	**Japanese**
4. Sino-Tibetan	**Chinese**
5. Uralo-Altaic	Estonian, Finnish, Hungarian, Azerki, **Turkish**, Uzbeki

The selected 'difficult' languages are in bold

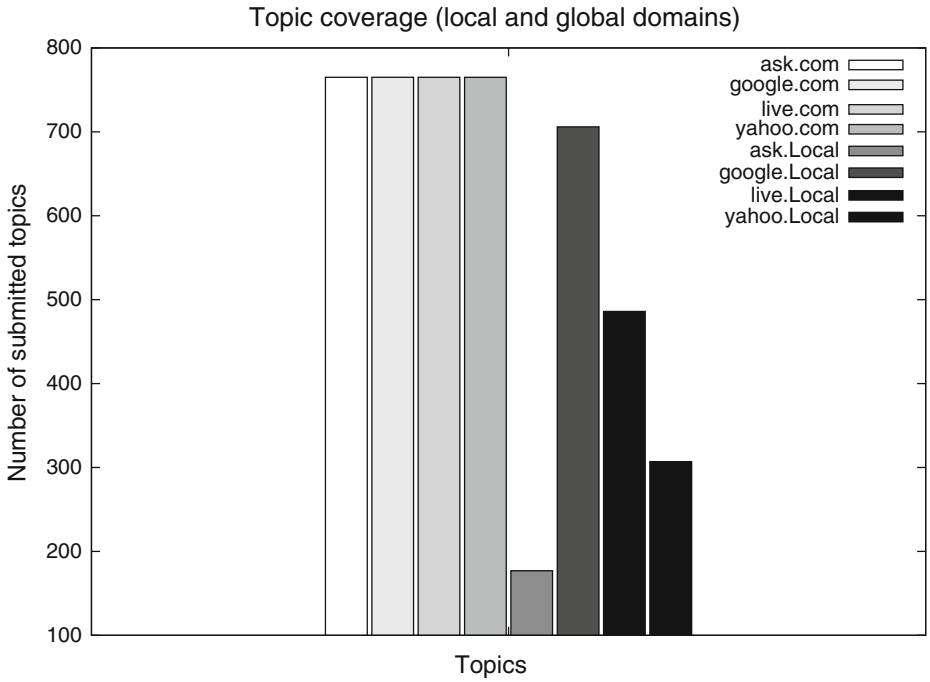

Fig. 1 Example of the scale of these experiments: number of queries submitted to the global and correct local domains only, resulting in over 4000 runs

Technology (NIST) implementation of Bernstein and Zobel's DECO algorithm (Bernstein and Zobel 2004), which was used in the TREC Terabyte tracks (Buttcher et al. 2006; Buttcher et al. 2005).

We used the *mean reciprocal rank* (MRR) within the top 10, which is a standard evaluation metric in this type of known-item search in TREC (Craswell et al. 2001) and Web CLEF (Balog et al. 2007). The reciprocal rank is calculated as one divided by the rank at which the (first and in this case only) relevant page is found. The mean reciprocal rank is obtained by averaging the reciprocal ranks of a set of topics (Balog et al. 2007). In all the results reported, we consider MRR = 0 if a homepage is not found in the first 10 results. In Sect. 5.2, we present the MRR scores, as well as histograms of ranks at which the correct answer appeared and proportions of not found homepages. We also present statistical testing in order to assess the differences in the rankings produced by the search engines when the features described in Sect. 4 are biasing the results: for repeated

measurements on a single sample, i.e., to compare the MRR of the same queries in different settings, we use the Wilcoxon matched-pairs signed-ranks test with p-values <0.05 (<0.01) denoting strong (respectively very strong) statistical significance; for testing if two populations, such as the rankings of the search engines, are different we use the χ^2 test (Owen and Jones 1986).

5.2 Evaluation results

We address our two research questions about the effect of language script and search domain upon retrieval as follows: Sect. 5.2.1 compares retrieval performance for queries in Latin versus non-Latin script. Section 5.2.2 compares retrieval performance for global versus local search domain. Section 5.2.3 combines the two previous points and compares retrieval performance for queries in Latin versus non-Latin script and for global versus local search domain. Section 5.2.4 compares retrieval performance for global versus local versus incorrect search domain.

5.2.1 Latin versus non-Latin script

Figure 2 plots the rank of the retrieved homepages (x-axis) against the % of homepages at a given rank (y-axis), separately for Latin and non-Latin script. This plot aggregates data from all the languages that use non-Latin script, and for all search engines and domains. The paired histogram distributions in Fig. 2 are statistically different, using the χ^2 test, for every pair of Latin versus non-Latin script, independently for every search engine, and using eleven levels (ranks 1 to 10, and not found). Note that a detailed break-up of retrieval precision per language with respect to Latin and non-Latin script (and also with respect to search domain) will be presented in Sect. 5.2.3, Table 3.

Figure 2 shows that notably more correct homepages are ranked top (1–3) when using the correct script (slightly above 55%) than when using the Latin script (slightly above 35%). For

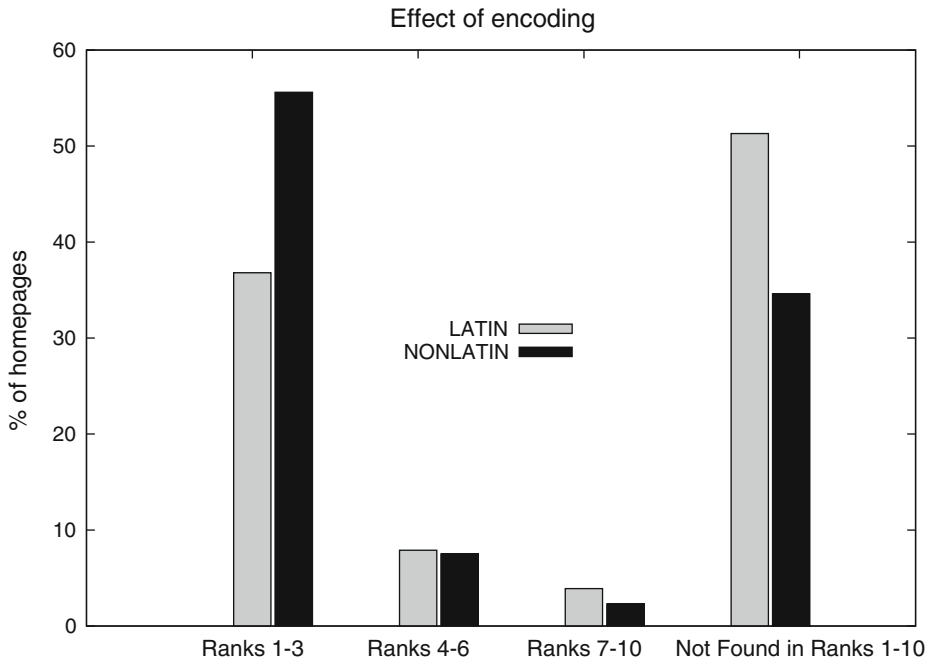

Fig. 2 Rank of the retrieved homepages (x-axis) against % of homepages at a given rank, for Latin and non-Latin script

Table 3 Best Mean Reciprocal Rank (MRR) scores for queries in their original script and latin script, for global and local search domains

Query language_script	Best global	Best local
Arabic_arabic	0.333 (google.com)	**0.492 (google.eg)**
Arabic_latin	0.354 (google.com)	0.432 (google.eg)
Armenian_armenian	**0.250 (yahoo.com)**	0.042 (google.am)
Armenian_latin	**0.250 (yahoo.com)**	0.042 (google.am)
Azerki_azerki	0.486 (yahoo.com)	**0.688 (google.az)**
Azerki_latin	0.569 (yahoo.com)	0.683 (google.az)
Bulgarian_bulgarian	0.321 (google.com)	**0.796 (google.bg)**
Bulgarian_latin	0.750 (google.com)	0.740 (google.bg)
Chinese_chinese	0.011 (google.com)	0.350 (google.cn)
Chinese_latin	**0.470 (google.com)**	0.392 (google.cn)
Greek_greek	0.411 (yahoo.com)	**0.953 (google.gr)**
Greek_latin	0.740 (google.com)	0.747 (google.gr)
Hebrew_hebrew	0.300 (ask.com)	0.520 (google.il)
Hebrew_latin	0.450 (google.com)	**0.733 (google.il)**
Iranian_iranian	0.062 (ask.com)	n/a
Iranian_latin	**0.384 (google.com)**	n/a
Japanese_japanese	0.750 (ask.com)	0.897 (google.jp)
Japanese_latin	**0.972 (google.com)**	**0.972 (google.jp)**
Russian_russian	0.438 (ask.com)	0.625 (yahoo.ru)
Russian_latin	0.677 (google.com)	**0.698 (yahoo.ru)**
Uzbeki_uzbeki	0.000 (google.com)	0.000 (google.uz)
Uzbeki_latin	**0.404 (google.com)**	0.375 (google.uz)

n/a Indicates that there is no available local domain. Statistical significance between local runs and their respective global runs is shown in Table 7

ranks 4–10, using the Latin script retrieves slightly more homepages than when using the correct script; this difference is marginal however. Most importantly, the percentage of homepages retrieved at ranks 4–10 is always <10%, notably lower than the corresponding percentage for ranks 1–3. In addition, the rate of homepages ranked 4–10 is always <10% regardless of the script used; this indicates that most homepages are retrieved either at ranks 1–3, or not retrieved. For ranks >10, which correspond to not-found homepages, the percentage is always ≥35%. When the Latin script is used, the rate of not-found homepages exceeds 50%, indicating that the correct script is less likely not to find homepages than the Latin script. The stark contrast between ranks 1–3 and ranks >10 indicates that:

- for queries written in the correct script of the language, more homepages are likely to be found than not-found; and most found homepages are likely to be in the top 3 ranks;
- for queries written in the Latin script, more homepages are likely to be not-found than found.

5.2.2 Global versus local search domain

Figure 3 plots the rank of the retrieved homepages (x-axis) against the % of homepages at a given rank (y-axis), for the global and local domains respectively. These plots aggregate data from all scripts and languages used, and present it separately for each search engine.

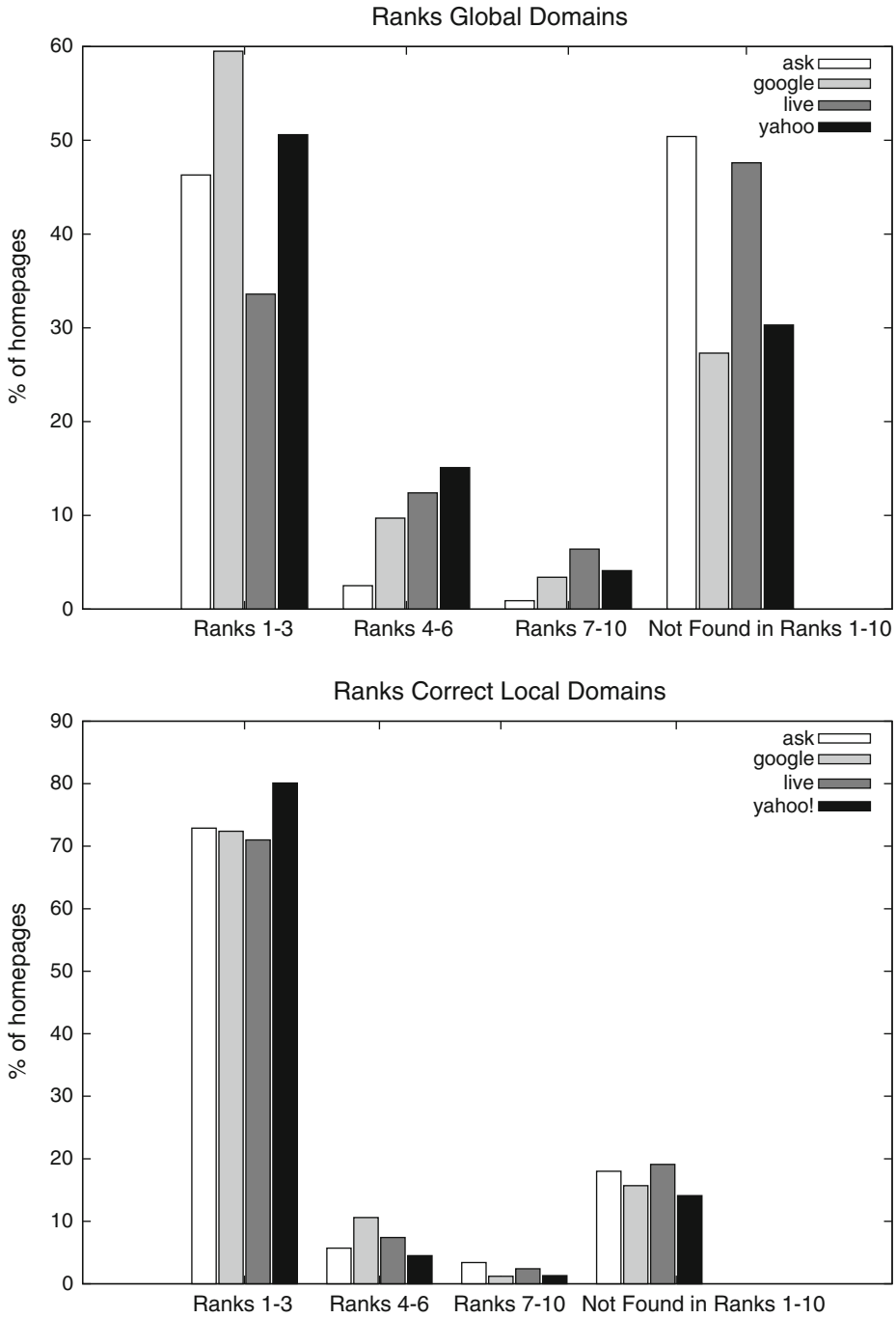

Fig. 3 Rank of the retrieved homepages (x-axis) against % of homepages at a given rank. Different bars represent different search engines respectively for the global domain (top) & local domain (bottom)

We use only languages with local domains available by more than one search engine. Similarly to Fig. 2, the paired histogram distributions in Fig. 3 are statistically different according to the χ^2 test, for every pair of local domain versus global domain, independently for every search engine, and using eleven levels (ranks 1 to 10, and not found).

We see that overall most of the homepages are retrieved at ranks 1–3 regardless of the domain and search engine used (the only exception being ask.com, where slightly more homepages are not found). In addition, similarly to the observations drawn from Fig. 2, there are notably more homepages at ranks 1–3 and not found than at ranks 4–10. Furthermore, we see that there are more homepages ranked 1–3 in the local domains than in

the global domains (this point is made explicit in the bottom plot of Fig. 3). Conversely, less homepages are found in the global domain than in the local domain. Recall that in Fig. 2, there was a similar relation between ranks and language script (the correct script performed better for the top ranks, and the Latin script had more not-found homepages). These observations indicate that, generally, refining the search by using the correct script or domain of a language helps to retrieve results and to rank them in the top 3.

5.2.3 Latin versus non-Latin script & global versus local search domain

Table 3 displays retrieval precision per language, with respect to Latin versus non-Latin script and global versus local domain. For this comparison, we use only languages of non-Latin script, and we present only the best MRR score among search engines. The best MRR per language is printed in bold. The results in Table 3 can be split into two: (i) languages scoring overall >0.500 (Azerki, Bulgarian, Greek, Hebrew, Japanese, Russian), and (ii) languages scoring overall <0.500 (Arabic, Armenian, Chinese, Iranian, Uzbeki). We discuss each category separately in the next two paragraphs.

In the >0.500 category, we see that the best runs use always the local domain (apart from Japanese, where best local = best global domain). Regarding script, half of these languages score better with non-Latin script, and the other half score better with Latin script, always in the local domain:

- For Azerki, Bulgarian, and Greek, the best scores are given by the non-Latin script and local domain combination. The scripts of these languages are interesting: Azerki[8] and Bulgarian use variants of the Cyrillic script (the main letters are the same, however the order of the letters may differ, and few characters may be added or omitted), and Greek uses the Greek script. Both Cyrillic and Greek are among the 'widespread' non-Latin scripts on the Web: users of Cyrillic are estimated at around 200 million; users of the Greek script are estimated at around 20 million speakers, however the script is also used widely for scientific notation worldwide.
- For Hebrew, Japanese, and Russian, the best scores are given by the local domain and Latin script combination (apart from Japanese as noted above). For Russian, the difference between the non-Latin and Latin script is rather small (0.625–0.698 respectively). However, for Hebrew and Japanese, the Latin script outperforms the non-Latin script notably (Hebrew: 0.733–0.520, Japanese: 0.972–0.897 respectively). Both of these languages use ideographic script, where there is not always a one-to-one correspondence to the Latin characters. In addition, the writing order differs from the common to most Indo-European languages top-down left-to-right order. Hebrew is written right-to-left, and Japanese is written right-to-left in top-down columns. In this light, the script of Hebrew and Japanese can be seen as more complex than the Cyrillic or Greek script, where the tokenisation of letters and writing order is standard Indo-European, and where the only thing that changes are the characters themselves. Hence, for search engines, the processing of Hebrew and Japanese script requires additional stages (compared to Cyrillic and Greek) in order to tokenise the symbols and address their reading order. This added processing may introduce noise to the search, especially because the automatic tokenisation of ideographic script and reading order normalisation are open problems in natural language processing.

[8] In the Republic of Azerbaijan, North Azerbaijani now officially uses the Latin alphabet, but the Cyrillic alphabet remains in wide use, while in Iran, South Azerbaijani still uses the Perso-Arabic script.

In the <0.500 category, we see that the best scores are generally given by global domain searches with Latin script (the only exception being Arabic where the best run uses Arabic script in the global domain; even in this case however, the difference between Latin and non-Latin script is small: 0.432–0.492 respectively). This general observation agrees with the conclusions drawn previously in Sects. 5.2.1–5.2.2, that the best performers (ranked 1–3) benefit more from the correct script and domain, while the worst performers (ranked >10) are associated with the Latin script and global domain. In this case, the languages scoring <0.500 MRR are our bad performers, and we see that they also tend to be more associated with the Latin script and global domain. The reason why these languages underperform with respect to the others could be due to their script. These languages use mainly ideographic script, which renders their automatic processing difficult, as discussed above. Specifically, in Arabic, Iranian, and Uzbeki (which use variants of the Arabic script), there is no capitalisation and no punctuation. Hence, whereas in most Indo-European languages, punctuation and capital letters denote the beginning of a sentence, in Arabic script, the reader must understand when a sentence finishes or begins. In addition, in Arabic script, words are combinations of letters which are attached to each other. However, some words are broken down because they contain letters which do not allow connections to be made to their end. One might assume that the current word has ended and the next word has started when in fact it is the same word. Furthermore, each letter looks different when it stands alone as a letter, when it is the first letter of a connected set of letters, when it is somewhere in the middle of a connection, and when it appears at the end of a set of connected letters. These features of Arabic script render its automatic processing rather difficult. Armenian uses its own script, which is classified as a branch of its own in the Indo-European languages (practically this means that there are no significant similarities to other alphabets). This script is not widely known in the world, partly because of the small number of native speakers (3 million speakers in Armenia and 8 million Armenians abroad), and partly because of changes in the script: in the 1920s, Soviet Armenia adopted a reformed spelling, which however was rejected by the Armenian diaspora (which out-numbers significantly the country's population). The result is that the already weak presence of Armenian on the Web (weak with respect to number of speakers) often lacks uniformity in script, which practically means noise for search engines. Finally, the last of the badly performing languages, Chinese, uses ideographic script, with non standard writing order, which as discussed above (in the case of Japanese) requires extra processing by the system. However, it should be noted that, unlike Armenian for instance, there is a very strong presence of Chinese script on the Web, hence the problem here is not so much with finding the data, but mainly with procesing it accurately.

Overall, the general conclusions drawn from Table 3 are that:

- For languages of 'easy' non-Latin script, the local domain and non-Latin script combination performs best. We describe as easy, non-Latin script that differs mainly from the Latin in characters, which has a one-to-one correspondence with Latin characters, and/or shares the same top-down left-to-right writing order.
- For languages of 'difficult' non-Latin script, the local domain with Latin script performs better. We describe as difficult, non-Latin script that is ideographic, and/or has a different writing order than Latin script languages.

Finally, Table 4 gives an overview of the results presented in Table 3 by showing the percentage of best MRR scores for each language script and search domain combination. We see that the best combinations are non-Latin script with local domain, and Latin script with global domain. The main message to take away from Table 4 is that, for non-Latin

Table 4 Percentage of how often the best MRR score (shown in Table 3) occurs with Latin or non-Latin script and local or global domain

	Latin script (%)	Non-latin script (%)
Local domain	25.0	**33.3**
Domain	**33.3**	8.3

Highest % in bold

script languages, using Latin script in local searches and also using non-Latin script in global searches underperform considerably.

5.2.4 Global versus local correct versus local incorrect search domain

Table 5 shows best MRR scores per language, domain, and search engine; (all MRR scores, not only the best, are shown in Table 7, in the Appendix). Best scores per language are in bold, and * (**) mark strong (very strong) statistically significant difference measured with the Wilcoxon signed ranks test for $p < 0.05$ ($p < 0.01$). We measure the statistical difference of any local domain run from its respective global domain run. We mark as n/a (not available) missing local domains. Overall, we see that querying the correct local domain gives the best MRR score on most occasions (54.1% of the times), followed by querying the incorrect local domain (35.1% of the times), and lastly the global domain (10.81% of the times). In fact, querying the global domain gives better MRR over any local domain only for two languages: Iranian (for which no search engine offers a local Iranian domain to our knowledge), and Swedish (for which the difference in MRR between the global and local domain is very small, namely 0.875–0.833 respectively). Regarding the incorrect domain, most of the queries submitted to it did not find any homepages. This result is also graphically presented in Fig. 4 as a histogram. Note that for some languages, best MRR scores are slightly higher for incorrect local domains than for the correct local domain, e.g., Arabic, Chinese, Estonian, Slovak, Uzbeki, Vietnamese. However, this observation holds *only for best MRR scores* among all search engines. There can be several reasons why best MRR scores are slightly higher for incorrect local domains than for the correct local domain. For instance, the name of the team may also be a place name which is well known within the country (hence ranked higher in the correct local domain) but not that well known outside the country (hence ranked lower in an incorrect local domain). Another more plausible reason could be that some engines (for instance Google) may detect the team as a *named entity* and display among the first 1–3 results short *news snippets* that are considered to be more relevant than the homepage itself; this does not happen so often when the query is sent to an incorrect domain. The final effect is that the real homepage is ranked lower in the result list, lowering MRR as well, despite of the fact that the search engine is doing a good job at recognising the query as a football team name.

Finally, Fig. 4 graphically summarises the overall performance of global versus local versus incorrect domain search for each search engine separately. Figure 4 shows the overall predominance in retrieval performance of the local domain, especially in the top ranks, and the weakness of the incorrect domain in finding homepages. There are some differences in the performance of each search engine, however the overall picture agrees with the general conclusion drawn from Sects. 5.2.1–5.2.3, namely that the best performing languages (the ones with most homepages ranked 1–3) benefit more from local

Table 5 Best Mean Reciprocal Rank (MRR) per language for global, correct, and incorrect domain

Language	Best global	Best local correct	Best local incorrect
Albanian	0.278 (google.com)	**0.283 (google.al)**	0.270 (google.cn)
Arabic	0.333 (google.com, ask.com)	0.492 (google.eg)	**0.556 (google.ru)**
Armenian	0.250 (yahoo.com)	0.042 (google.am)	**0.565 (google.il)**
Azerki	0.486 (yahoo.com)	**0.683 (google.az)**	0.625 (yahoo.ru)
Bosnian	0.411 (google.com)	**0.519 (google.ba)**	0.362* (google.tr)
Bulgarian	0.473 (live.com)	**0.796 (google.bg)**	0.451 (google.tr)
Chinese	0.470 (google.com)	0.392 (google.cn)	**0.565 (google.il)**
Croatian	0.621 (google.com)	**0.864** (google.cr)**	0.818** (google.tr)
Czech	0.700 (google.com)	0.733** (google.cz)	**0.800** (google.tr)**
Danish	0.848 (google.com)	**0.894** (google.dk)**	0.758** (google.tr)
Dutch	0.667 (ask.com)	**0.781** (yahoo.nl)**	0.622 (google.tr)
English	0.938 (google.com)	**1.000** (live.au)**	**1.000 (google.il)**
Estonian	0.319 (google.com)	0.636** (google.et)	**0.783* (yahoo.ru)**
Finnish	0.786 (ask.com)	**0.893** (yahoo.fi)**	0.810** (live.ru)
French	0.658 (ask.com)	0.730** (yahoo.fr)	**0.732 (google.tr)**
German	**0.787 (ask.com)**	0.746 (ask.de)	0.750 (google.il)
Greek	0.510 (live.com)	**0.953** (google.gr)**	0.688 (google.tr)
Hebrew	0.469 (yahoo.com)	**0.733 (google.il)**	0.514 (yahoo.ru)
Hungarian	0.391 (google.com)	**0.721** (google.hu)**	0.412* (google.il)
Iranian	**0.384 (google.com)**	–	0.000 (all)
Italian	0.625 (ask.com)	**0.842** (yahoo.fi)**	0.459** (google.il)
Japanese	**0.972 (google.com)**	**0.972 (google.jp)**	0.720 (live.ru)
Lithuanian	0.512 (google.com)	**0.786 (google.lt)**	0.325** (live.cn)
Polish	0.639 (google.com)	0.535** (google.pl)	**0.708** (google.il)**
Portuguese	0.759 (ask.com)	**0.828** (google.pt)**	0.711** (google.il)
Romanian	0.576 (google.com)	0.706** (google.ro)	**0.750** (google.il)**
Russian	0.677 (google.com)	**0.698 (yahoo.ru)**	0.565 (google.il)
Slovak	0.669 (google.com)	0.750** (google.sk)	**0.771** (google.il)**
Slovenian	0.537 (google.com)	**0.759* (google.si)**	0.722* (google.jp)
Spanish	0.950 (ask.com)	**1.000** (yahoo.es)**	**1.000** (google.ru)**
Swedish	**0.875 (ask.com)**	0.833** (google.se)	0.797 (google.tr)
Turkish	0.741 (ask.com)	**0.917** (google.tr)**	0.850** (yahoo.ru)
Uzbeki	0.404 (google.com)	0.375 (google.uz)	**0.565 (google.il)**
Vietnamese	0.292 (ask.com)	0.243 (google.vn)	**0.385 (yahoo.ru)**
% best	10.81%	54.1%	35.1%

For languages with more than one domain, e.g., Spanish, we show the domain of the best MRR

Best MRR per language is in bold

– Indicates that there is no available domain

* (**) Indicates strong (very strong) statistical significance of any local domain from its respective global domain, with $p < 0.05$ (<0.01) using the Wilcoxon signed-rank test

domain searches, and that the worst performing languages (the ones with the most not found homepages) are associated with non-local domain searches. Regarding worst performers, in this section we see that local searches are outperformed not only by global

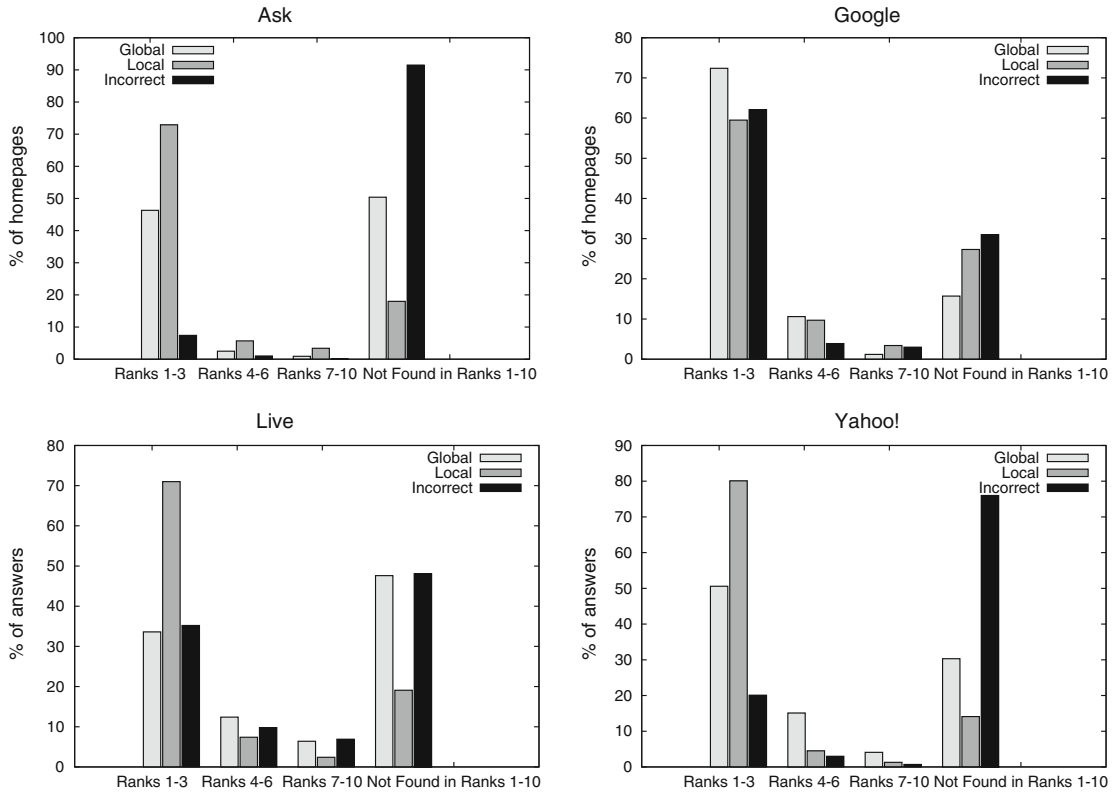

Fig. 4 Retrieval performance with global, local and incorrect domain for each search engine

domain searches, but also by incorrect domain searches. This may be an indication that these queries are particularly hard, and that further language-specific processing may be required by the system.

In addition, the MRR scores in Table 5 are at least comparable to the best MRR scores of the named page finding task of the WebCLEF mixed monolingual track and the TREC Terabyte track for the English language only, and even though this task is much harder (as described in Sect. 2). For 2006, the best mixed monolingual WebCLEF MRR score was 0.531 (Balog et al. 2007), and for 2005 and 2006, the best Terabyte MRR scores for named page finding were 0.463 and 0.512 respectively (Buttcher et al. 2006; Buttcher et al. 2005), while the average MRR score of all languages used in our experiments (averaged over the best MRR score for each language) is 0.734. This shows that state of the art search engine performance is very satisfactory for most languages. An exception to this are very few languages, for which there is still room for improvement, e.g., Iranian (best MRR: 0.384 and no local search domain available), Albanian and Vietnamese (best MRRs: 0.283 and 0.292 respectively, local search domains only available by google).

Table 6 contains the percentage of homepages that were not found in the top 10 ranks, for each local domain. For comparison, note that the TREC Terabyte track corresponding percentages of not found Web pages for English only were 17.1% in 2005 (Buttcher et al. 2005) and 12.7% in 2006 (Buttcher et al. 2006). Overall, we observe that the majority of the percentages we report are less than 20%: out of 71 domains, 44 domains have $<$20% not found homepages, and 28 domains have $>$20% not found homepages. This performance is encouraging, considering that we query the whole Web, as opposed to a static crawl of the Web only as in the case of the Terabyte named page finding task. In most cases very few homepages are not found in the top 10. There are few domains for which 50% or more of the homepages are not found, which may be due to several reasons: for instance, for some domains there were very few

Table 6 Percentage of not found homepages (in the top 10) per local search domain

Local domain	Not found (%)
Australia, Bahrain, Brazil, Cayman Islands Cote d'Ivoire, Denmark, Ecuador, France, Ireland, Israel, Japan Latvia, Luxembourg, Malta, Portugal, Spain, Sudan, USA	0
Italy	5
Greece, Hungary, Northern Ireland, Sweden, Turkey	6
Bulgaria, Finland	7
South Africa	8
Belarus, Croatia	9
Austria	10
Bosnia-Herzegovina, Slovenia	11
Russia	12
Lithuania	14
Venezuela	15
Azerbaijan, Belgium, Bolivia, Egypt, Mexico, Netherlands, Slovakia	17
Poland	19
Argentina, Canada, Costa Rica, Czech Republic, Wales	20
Chile, Germany	22
Colombia, Lebanon	25
El Salvador, Romania	29
China, England	30
Albania	33
Estonia	36
Iran	38
Cyprus, Jordan	40
Scotland, Syria, Uzbekistan, Vietnam	50
Algeria	60
Zimbabwe	67
Armenia	75
Aruba, Belize, Cameroon	100

Results obtained using google local sites if applicable (otherwise, google.com)

queries, hence not finding even one page could lower significantly the overall percentage, e.g., Syria: 6 queries, Uzbekistan: 4 queries, Zimbabwe: 3 queries, Armenia: 4 queries, Aruba: 1 query, Belize: 1 query, Cameroon: 1 query. Furthermore, for other domains, like Scotland, there was no local search domain, so a global domain was used; however several Scottish team names are quite common as team names worldwide (e.g., Celtic, Rangers), making it harder to retrieve Scottish homepages from the global domain.

6 Conclusions

The influx of tens of millions of new Web users each year can be expected to have far-reaching consequences for universal digital information access. At the very least, the Web

will offer ever greater numbers of multilingual and much more sophisticated information and communications. Search engines need to provide fair and equal access to information, regardless of the language in which a query is written or where the query was posted from. In this work, we asked two questions: How do existing state of the art search engines deal with languages written in different alphabets (scripts)? Do local language-based search domains actually facilitate access to information? We restricted our study to the very realistic retrieval scenario of users seeking homepages of football teams. Experiments with four major commercial and freely available search engines, which we used to query the whole Web in 34 languages, 71 local domains, and by submitting 764 queries in more than 10,000 runs, showed that:

- Queries issued in the correct script of a language are more likely to be found and ranked in the top 3, while queries in non-Latin script languages issued in Latin script are less likely to be found. However, for particularly difficult non-Latin scripts, such as those using ideographic writing of right-to-left reading order, the Latin script searches outperform non-Latin script searches. This indicates that the search engines need to provide advanced language-specific processing for the languages, apart from support for their encoding.
- Queries issued to the correct local domain of a search engine, e.g., Russian queries to yahoo.ru, are likely to have better retrieval performance than queries issued to the global domain of a search engine, for non-Latin script languages; for Latin script languages, global domain searches outperform local domain searches. Note that local search domains are missing for several languages, which is an area of potential improvement.
- Queries issued to the incorrect local domain of a search engine, e.g., Russian queries to yahoo.es, are less likely to retrieve any homepages than queries issued to the local or global domain of a search engine, for most languages. We can report a few exceptions to this, where for some particularly hard languages, the incorrect domain performs better than the correct local or global domain. This is an indication that search should improve for these languages, by employing more language-specific processing.

The general conclusion emerging from this study is that the industry is investing heavily in non-English Web retrieval, through language-specific portals and character support, among other things, and that major commercial and freely available search engines offer high quality retrieval performance (as evaluated with respect to reported WebCLEF and TREC performances). It is encouraging to see good performance for languages that are both 'difficult' and also under-represented on the Web (e.g., best MRR for Azerki is 0.683, for Greek is 0.953, for Hebrew is 0.733), because this raises significantly the baseline for all other languages.

Future research directions include refining the way in which we detect duplicate mirrored or redirected homepages, as mentioned in Sect. 5.1, as well as contextualising the queries (for instance by inserting the word 'football') in order to overcome cases where the query name is identical to a proper noun, such as a city. It would also be interesting to test some 'truly' local search engines which have a large market share in their countries, e.g., Baidu (http://www.baidu.com/) for China, and compare the results to the mainstream search engines used in this study.

Acknowledgements Roi Blanco is co-funded by FEDER, Ministerio de Ciencia e Innovación and Xunta de Galicia under projects TIN2008-06566-C04-04 and 07SIN005206PR. Christina Lioma is funded by K.U.L. Postdoctoral Fellowship F+/08/002.

Appendix

Table 7 Best Mean Reciprocal Rank (MRR) per language for global, correct, and incorrect domain

Domain	Albanian	Arabic	Armenian	Azerki	Bosnian	Bulgarian
Global	ask.com: 0.000	ask.com: 0.333	ask.com: 0.000	ask.com: 0.167	ask.com: 0.192	ask.com: 0.296
	google.com: 0.278	google.com: 0.333	google.com: 0.042	google.com: 0.228	google.com: 0.411	google.com: 0.321
	live.com: 0.056	live.com: 0.091	live.com: 0.000	live.com: 0.385	live.com: 0.222	live.com: 0.473
	yahoo.com: 0.241	yahoo.com: 0.306	yahoo.com: 0.250	yahoo.com: 0.486	yahoo.com: 0.273	yahoo.com: 0.246
Local	**google.al: 0.283**	google.eg: 0.492	google.am: 0.042	**google.az: 0.683**	**google.ba**: 0.519**	**google.bg: 0.796**
Incorrect	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 000	ask.jp: 0.000
	google.cn: 0.270	**google.ru: 0.556**	**google.il: 0.565**	google.jp: 0.226	google.tr: 0.362*	google.tr: 0.451
	live.cn: 0.015	live.ru: 0.361	live.cn: 0.125	live.ru: 0.394	live.cn: 0.127	live.cn: 0.152
	yahoo.jp: 0.000	yahoo.ru: 0.500	yahoo.jp: 0.000	yahoo.ru: 0.625	yahoo.cn: 0.136*	yahoo.cn: 0.153

Domain	Chinese	Croatian	Czech	Danish	Dutch	English
Global	ask.com: 0.100	ask.com: 0.545	ask.com: 0.697	ask.com: 0.530	ask.com: 0.667	ask.com: 0.312
	google.com: 0.470	google.com: 0.621	google.com: 0.700	google.com: 0.848	google.com: 0.431	google.com: 0.938
	live.com: 0.352	live.com: 0.175	live.com: 0.275	live.com: 0.189	live.com: 0.303	live.com: 0.500
	yahoo.com: 0.116	yahoo.com: 0.364	yahoo.com: 0.383	yahoo.com: 0.388	yahoo.com: 0.378	yahoo.com: 0.812
Local	–	–	–	–	ask.nl: 0.673	ask.uk: 0.704
	google.cn: 0.392	**google.cr: 0.864****	google.cz: 0.733**	**google.dk: 0.894****	google.nl: 0.520**	google.au: 0.938
	live.cn: 0.300	–	–	live.dk: 0.712**	live.nl: 0.740**	**live.au: 1.000****
	yahoo.cn: 0.000	–	–	yahoo.dk: 0.768**	**yahoo.nl: 0.781****	yahoo.au: 0.938**
Incorrect	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.056**	ask.jp: 0.000
	google.il: 0.565	google.tr: 0.818**	**google.tr: 0.800****	google.tr: 0.758**	google.tr: 0.622	**google.il: 1.000**
	live.ru: 0.125	live.cn: 0.393**	live.cn: 0.370**	live.cn: 0.405**	live.cn: 0.495**	live.cn: 0.424**
	yahoo.jp: 0.000	yahoo.cn: 0.464**	yahoo.cn: 0.506**	yahoo.cn: 0.333*	yahoo.cn: 0.583**	yahoo.cn: 0.750**

Table 7 continued

Domain	Estonian	Finnish	French	German	Greek	Hebrew
Global	ask.com: 0.000	ask.com: 0.786	ask.com: 0.658	**ask.com: 0.787**	ask.com: 0.312	ask.com: 0.150
	google.com: 0.319	google.com: 0.419	google.com: 0.481	google.com: 0.655	google.com: 0.383	google.com: 0.450
	live.com: 0.094	live.com: 0.180	live.com: 0.168	live.com: 0.095	live.com: 0.510	live.com: 0.210
	yahoo.com: 0.212	yahoo.com: 0.378	yahoo.com: 0.379	yahoo.com: 0.318	yahoo.com: 0.411	yahoo.com: 0.469
Local	–	–	ask.fr: 0.618	ask.de: 0.746	–	–
	google.et: 0.636**	google.fi: 0.881**	google.fr: 0.542**	google.de: 0.736*	**google.gr: 0.953**	**google.il: 0.733**
	–	live.fi: 0.776**	–	live.de: 0.687**	–	–
	–	**yahoo.fi: 0.893****	yahoo.fr: 0.730**	yahoo.de: 0.587**	–	–
Incorrect	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.238	ask.jp: 0.000	ask.jp: 0.000
	google.il: 0.615**	google.cn: 0.607**	**google.tr: 0.732**	google.il: 0.750*	google.tr: 0.688	google.jp: 0.143
	live.ru: 0.503	live.ru: 0.810**	live.cn: 0.463**	live.jp: 0.067	live.cn: 0.481	live.ru: 0.124
	yahoo.ru: 0.783*	yahoo.cn: 0.548**	yahoo.cn: 0.374	yahoo.cn: 0.536**	yahoo.cn: 0.290	yahoo.ru: 0.514
Domain	Hungarian	Iranian	Italian	Japanese	Lithuanian	Polish
Global	ask.com: 0.156	ask.com: 0.062	ask.com: 0.625	ask.com: 0.269	ask.com: 0.190	ask.com: 0.408
	google.com: 0.391	**google.com: 0.384**	google.com: 0.176	**google.com: 0.972**	google.com: 0.512	google.com: 0.639
	live.com: 0.040	live.com: 0.268	live.com: 0.095	live.com: 0.928	live.com: 0.024	live.com: 0.418
	yahoo.com: 0.208	yahoo.com: 0.284	yahoo.com: 0.293	yahoo.com: 0.788	yahoo.com: 0.207	yahoo.com: 0.341
Local	–	–	ask.it: 0.494**	ask.jp: 0.944	–	–
	google.hu: 0.721**	–	google.it: 0.517**	**google.jp: 0.972**	**google.lt: 0.786**	google.pl: 0.535**
	–	–	live.it: 0.583**	–	–	–
	–	–	**yahoo.it: 0.842****	yahoo.jp: 0.950	–	–
Incorrect	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.013**	ask.ru: 0.000	ask.jp: 0.000	ask.jp: 0.000
	google.il: 0.412*	google.il: 0.000	google.il: 0.459**	google.cn: 0.685	google.ru: 0.167**	**google.il: 0.708****
	live.jp: 0.212**	live.cn: 0.000	live.jp: 0.067	live.ru: 0.720	live.cn: 0.325**	live.cn: 0.135
	yahoo.cn: 0.058**	yahoo.jp: 0.00	yahoo.cn: 0.365*	yahoo.cn: 0.597	yahoo.jp: 0.143	yahoo.cn: 0.281**

Table 7 continued

Domain	Portuguese	Romanian	Russian	Slovak	Slovenian	Spanish
Global	ask.com: 0.759	ask.com: 0.412	ask.com: 0.396	ask.com: 0.667	ask.com: 0.111	ask.com: 0.950
	google.com: 0.549	google.com: 0.576	google.com: 0.677	google.com: 0.669	google.com: 0.537	google.com: 0.533
	live.com: 0.327	live.com: 0.291	live.com: 0.450	live.com: 0.156	live.com: 0.130	live.com: 0.318
	yahoo.com: 0.409	yahoo.com: 0.148	yahoo.com: 0.590	yahoo.com: 0.381	yahoo.com: 0.158	yahoo.com: 0.517
Local	–	–	–	–	–	ask.es: 0.950
	google.pt: 0.828**	google.ro: 0.706**	google.ru: 0.609	google.sk: 0.750**	**google.si: 0.759***	google.es: 0.838***
	live.pt: 0.692**	–	live.ru: 0.694	–	–	live.es: 0.823**
	–	–	**yahoo.ru: 0.698**	–	–	**yahoo.es: 1.000****
Incorrect	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000
	google.il: 0.711*	**google.il: 0.750****	google.il: 0.565	**google.il: 0.771****	google.ru: 0.167**	google.il: 0.708**
	live.cn: 0.214**	live.cn: 0.207	live.cn: 0.125	live.cn: 0.324	live.cn: 0.325**	live.cn: 0.135
	yahoo.cn: 0.323**	yahoo.jp: 0.111**	yahoo.jp: 0.000	yahoo.cn: 0.378	yahoo.jp: 0.143	yahoo.cn: 0.281**

Domain	Swedish	Turkish	Uzbeki	Vietnamese
Global	**ask.com: 0.875**	ask.com: 0.741	ask.com: 0.250	ask.com: 0.292
	google.com: 0.773	google.com: 0.715	google.com: 0.404	google.com: 0.179
	live.com: 0.159	live.com: 0.380	live.com: 0.312	live.com: 0.170
	yahoo.com: 0.459	yahoo.com: 0.479	yahoo.com: 0.369	yahoo.com: 0.050
Local	–	–	–	–
	google.se: 0.833**	**google.tr: 0.917****	google.uz: 0.375	google.vn: 0.243
	live.se: 0.740**	live.tr: 0.620**	–	–
	yahoo.se: 0.807**	–	–	–

Table 7 continued

Domain	Swedish	Turkish	Uzbeki	Vietnamese
Incorrect	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000	ask.jp: 0.000
	google.tr: 0.797	google.il: 0.815**	**google.il: 0.565**	google.il: 0.018**
	live.jp: 0.387**	live.ru: 0.683**	live.cn: 0.125	live.jp: 0.139
	yahoo.cn: 0.341**	yahoo.ru: 0.850**	yahoo.jp: 0.000	**yahoo.ru: 0.385**

For languages with more than one domain, e.g., Spanish, we show the domain of the best MRR

Best MRR per language is in bold

– Indicates that there is no available domain

* (**) Indicates strong (very strong) statistical significance of any local domain from its respective global domain, with $p < 0.05$ (<0.01) using the Wilcoxon signed-rank test

References

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., & Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. *ACM Transaction on Asian Language Information Processing, 6*(4), 1–33.

Adriani, M., & Pandugita, R. (2006). Using the Web information structure for retrieving Web pages. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, et al. (Eds.), Accessing multilingual information repositories. *6th Workshop of the Cross-Language Evalution Forum, CLEF 2005*, Vienna, Austria, 21–23 September, 2005. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4022, pp. 892–897). Heidelberg: Springer.

Ahmed, F., & Nurnberger, A. (2007). N-grams conflation approach for Arabic text. In F. Lazarinis, J. V. Ferro, & J. Tait (Eds.), *Proceedings of the Workshop on Improving Web Retrieval for Non-English Queries (iNEWS), SIGIR 2007*, Amsterdam, The Netherlands, 23–27 July, 2007.

Balog, K., Azzopardi, L., Kamps, J., & de Rijke, M. (2007). Overview of WebCLEF 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 803–819). Heidelberg: Springer.

Balog, K., & de Rijke, M. (2007). Index combinations and query reformulations for mixed monolingual Web retrieval. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 830–833). Heidelberg: Springer.

Berners-Lee, T. (2005). WWW at 15 years: Looking forward. In A. Ellis & T. Hagino (Eds.), *WWW*, p. 1. New York: ACM.

Bernstein, Y., & Zobel, J. (2004). A scalable system for identifying co-derivative documents. In A. Apostolico & M. Melucci (Eds.), SPIRE. *Lecture Notes in Computer Science* (Vol. 3246, pp. 55–67). Heidelberg: Springer.

Broder, A. Z. (2002). A taxonomy of Web search. *SIGIR Forum, 36*(2), 3–10.

Buttcher, S., Clarke, C., & Soboroff, I. (2006). TREC Terabyte track overview. In *TREC*. Gaithersburg: NIST.

Buttcher, S., Scholer, F., & Soboroff, I. (2005). TREC Terabyte track overview. In *TREC*. Gaithersburg: NIST.

Chung, W. (2008). Web searching in a multilingual world. In *Communications of the ACM* (Vol. 51, pp. 32–40). New York: ACM.

Craswell, N., Hawking, D., & Robertson, S. E. (2001). Effective site finding using link anchor information. In W. B. Croft, D. J. Harper, D. H. Kraft, & J. Zobel (Eds.), *SIGIR* (pp. 250–257). New York: ACM.

Craswell, N., Hawking, D., Wilkinson, R., & Wu, M. (2001). TREC10 Web and interactive tracks at CSIRO. In *TREC*. Gaithersburg: NIST.

Crystal, D. (2006). *Language and the Internet*. Cambridge University Press: New York.

Daya, E., Roth, D., & Wintner, S. (2008). Identifying semitic roots: Machine learning with linguistics constraints. *Computational Linguistics, 34*(3), 429–448.

Efthimiadis, E., Malevris, N., Kousaridas, A., Lepeniotou, A., & Loutas, N. (2007). How do search engines handle Greek queries? In F. Lazarinis, J. V. Ferro, & J. Tait (Eds.), *Proceedings of the Workshop on Improving Web Retrieval for Non-English Queries (iNEWS), SIGIR 2007*, Amsterdam, The Netherlands, 23–27 July, 2007.

Fallows, D. (2007). China's online population explosion. *Pew Internet & American life project*. http://www.pewinternet.org/ppf/r/218/report_display.asp.

Figuerola, C. G., Berrocal, J. L. A., Rodríguez, Á. F. Z., & de Aldana, E. R. V. (2006). Web page retrieval by combining evidence. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, et al. (Eds.), Accessing multilingual information repositories. *6th Workshop of the Cross-Language Evalution Forum, CLEF 2005*, Vienna, Austria, 21–23 September, 2005. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4022, pp. 880–887). Springer.

Garfield, R. (1975). Transliteration, transcription, translation. *Essays of an Information Scientist, 2*(16), 254–256.

Gey, F. C., Kando, N., Lin, C. Y., & Peters, C. (2006). New directions in multilingual information access. *SIGIR Forum, 40*(2), 31–39.

Hasan, M., & Matsumoto, Y. (2000). Chinese-Japanese cross language information retrieval: A Han character based approach. In *ACL Workshop on Word Senses and Multi-linguality* (pp. 19–26). Hong Kong: ACL.

Heuwing, B., Mandl, T., & Strötgen, R. (2007). Multilingual Web retrieval experiments with field specific indexing strategies for WebCLEF 2006 at the University of Hildesheim. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante,

Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 834–837). Heidelberg: Springer.

Jensen, N., Hackl, R., Mandl, T., & Strötgen, R. (2006). Web retrieval experiments with the EuroGOV corpus at the University of Hildesheim. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, et al. (Eds.), Accessing multilingual information repositories. *6th Workshop of the Cross-Language Evalution Forum, CLEF 2005*, Vienna, Austria, 21–23 September, 2005. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4022, pp. 837–845). Heidelberg: Springer.

Kamps, J., de Rijke, M., & Sigurbjörnsson, B. (2006). Combination methods for crosslingual Web retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, et al. (Eds.), Accessing multilingual information repositories. *6th Workshop of the Cross-Language Evalution Forum, CLEF 2005*, Vienna, Austria, 21–23 September, 2005. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4022, pp. 856–864). Heidelberg: Springer.

Lazarinis, F. (2007). Web retrieval systems and the Greek language: Do they have an understanding?. *Journal of Information Science, 33*(5), 622–636.

Lazarinis, F., Ferro, J. V., & Tait, J. (2007). Improving non-English Web searching (iNEWS07). *SIGIR Forum, 41*(2), 72–76.

López, F. R., Jiménez-Salazar, H., & Pinto, D. (2007). Vocabulary reduction and text enrichment at WebCLEF. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 838–843). Heidelberg: Springer.

Macdonald, C., Lioma, C., & Ounis, I. (2007). Terrier takes on the non-English Web. In F. Lazarinis, J. V. Ferro, & J. Tait (Eds.), *Proceedings of the Workshop on Improving Web Retrieval for Non-English Queries (iNEWS), SIGIR 2007*, Amsterdam, The Netherlands, 23–27 July, 2007.

Macdonald, C., Plachouras, V., He, B., Lioma, C., & Ounis, I. (2006). University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, et al. (Eds.), Accessing multilingual information repositories. *6th Workshop of the Cross-Language Evalution Forum, CLEF 2005*, Vienna, Austria, 21–23 September, 2005. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4022, pp. 898–907). Heidelberg: Springer.

Martínez-González, Á., Martínez-Fernández, J. L., de Pablo-Sánchez, C., & Villena-Román, J. (2006). Miracle at WebCLEF 2005: Combining Web specific and linguistic information. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, et al. (Eds.), Accessing multilingual information repositories. *6th Workshop of the Cross-Language Evalution Forum, CLEF 2005*, Vienna, Austria, 21–23 September, 2005. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4022, pp. 869–872). Heidelberg: Springer.

McNamee, P. (2007). JHU/APL ad hoc experiments at CLEF 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 157–162). Heidelberg: Springer.

Mishne, G. (2007). *Applied text analytics for blogs*. PhD in Computing Science, University of Amsterdam.

Orengo, V. M., Buriol, L. S., & Coelho, A. R. (2007). A study on the use of stemming for monolingual ad-hoc Portuguese information retrieval. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 91–98). Heidelberg: Springer.

Owen, F., & Jones, R. (1986). *Statistics*. Pitman, London.

Pinto, D., Rosso, P., & Jiménez, E. (2007). A penalisation-based ranking approach for the mixed monolingual task of WebCLEF 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 826–829). Heidelberg: Springer.

Rodríguez, Á. F. Z., Berrocal, J. L. A., & Figuerola, C. G. (2007). Local query expansion using terms windows for robust retrieval. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 145–152). Heidelberg: Springer.

Santiago, F. M., Ráez, A. M., Cumbreras, M. A. G., & López, L. A. U. (2007). SINAI at CLEF 2006 ad hoc robust multilingual track: Query expansion using the Google search engine. In C. Peters, P. Clough,

F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 119–126). Heidelberg: Springer.

Savoy, J., & Abdou, S. (2007). Experiments with monolingual, bilingual, and robust retrieval. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 137–144). Heidelberg: Springer.

Sigurbjörnsson, B., Kamps, J., & de Rijke, M. (2006). EuroGOV. Engineering a multilingual Web corpus. In C. Peters, F. Gey, J. Gonzalo, G. Jones, M. Kluck, B. Magnini, et al. (Eds.), Accessing multilingual information repositories. *6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, *Lecture Notes in Computer Science* (Vol. 4022, pp. 825–836). Heidelberg: Springer.

Tomlinson, S. (2006). Danish & Greek Web search experiments with Hummingbird Searchserver^{tm} at CLEF 2005. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, et al. (Eds.), Accessing multilingual information repositories. *6th Workshop of the Cross-Language Evalution Forum, CLEF 2005*, Vienna, Austria, 21–23 September, 2005. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4022, pp. 846–855). Heidelberg: Springer.

Tomlinson, S. (2007). Comparing the robustness of expansion techniques and retrieval measures. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 129–136). Heidelberg: Springer.

Tzekou, P., Stamou, S., Zotos, N., & Kozanitis, L. (2007). Querying the Greek Web in Greeklish. In F. Lazarinis, J. V. Ferro, & J. Tait (Eds.), *Proceedings of the Workshop on Improving Web Retrieval for Non-English Queries (iNEWS), SIGIR 2007*, Amsterdam, The Netherlands, 23–27 July, 2007.

Van Rijsbergen, C. J. K. (1979). *Information retrieval*. London: Butterworth.

Vilares, J., Oakes, M. P., & Tait, J. (2007). A first approach to CLIR using character-grams alignment. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 111–118). Heidelberg: Springer.

Voegelin, C. F., & Voegelin, F. M. (1977). *Classification and index of the world's languages*. New York: Elsevier.

Voorhees, E. M., & Harman, D. K. (2005). *TREC: Experiment and evaluation in information retrieval*. Cambridge, MA: MIT Press.

Wijaya, S., Widhi, B., Khoerniawan, T., & Adriani, M. (2007). Applying relevance feedback for retrieving Web-page retrieval. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), Evaluation of multilingual and multi-modal information retrieval. *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, 20–22 September, 2006. Revised Selected Papers, *Lecture Notes in Computer Science* (Vol. 4730, pp. 848–851). Heidelberg: Springer.