Arabic Information Retrieval

Kareem Darwish
Qatar Computing Research Institute
kdarwish@qf.org.qa

Walid Magdy
Qatar Computing Research Institute
wmagdy@qf.org.qa

Full text available at: http://dx.doi.org/10.1561/1500000031
Full text available at: http://dx.doi.org/10.1561/1500000031
Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2013, Volume 7, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Full text available at: http://dx.doi.org/10.1561/1500000031
Arabic Information Retrieval

Kareem Darwish  
Qatar Computing Research Institute  
kdarwish@qf.org.qa

Walid Magdy  
Qatar Computing Research Institute  
wmagdy@qf.org.qa
Contents

1 Introduction 2
  1.1 The Arabic Language 3
  1.2 The Remainder of the Survey 6

2 Arabic Features Affecting Retrieval 7
  2.1 Arabic Orthography and Print 7
  2.2 Arabic Morphology 9
  2.3 Arabic Dialects 13
  2.4 Arabizi 15
  2.5 Arabic Speech 17
  2.6 Arabic on the Web 19

3 Arabic Preprocessing and Indexing 21
  3.1 Handling Encodings and Transliteration Schemes 21
  3.2 Handling Orthography 23
  3.3 Handling Morphology 28
  3.4 Handling Stopwords 31
  3.5 Handling Spelling and Lexical Choice Variations 31
  3.6 Best Index Terms 32
  3.7 Retrieval Models 34
Abstract

In the past several years, Arabic Information Retrieval (IR) has garnered significant attention. The main research interests have focused on retrieval of formal language, mostly in the news domain, with ad hoc retrieval, OCR document retrieval, and cross-language retrieval. The literature on other aspects of retrieval continues to be sparse or non-existent, though some of these aspects have been investigated by industry. Others aspects of Arabic retrieval that have received attention include document image retrieval, speech search, social media and web search, and filtering. However, efforts on different aspects of Arabic retrieval continue to be deficient and severely lacking behind efforts in other languages. The survey covers: 1) general properties of the Arabic language; 2) some of the aspects of Arabic that affect retrieval; 3) Arabic processing necessary for effective Arabic retrieval; 4) Arabic retrieval in public IR evaluations; 5) specialized retrieval problems, namely Arabic-English CLIR, Arabic Document Image Retrieval, Arabic Social Search, Arabic Web Search, Question Answering, Image retrieval, and Arabic Speech Search; 6) Arabic IR and NLP resources; and 7) open IR problems that require further attention.

DOI: 10.1561/1500000031.
Most early studies on Arabic IR relied on relatively small test collections containing hundreds of documents that are composed of character-coded Arabic text \cite{7,13,88}. Increased interest in Arabic processing and retrieval in the early 2000’s led to significant work that mostly relied on a single large collection (from TREC-2001/2002) \cite{68,129}. However, most of the work was restricted to ad hoc retrieval and cross-language retrieval. Later work focused on other aspects of Arabic retrieval including document image retrieval, speech search, social media and web search, and filtering. However, efforts on different aspects of Arabic retrieval continue to be deficient and severely lacking behind efforts in other languages. This survey reviews recent literature pertaining to different aspects of Arabic IR including different domains and applications. It also describes some of the Arabic specific challenges affecting retrieval and some of the proposed solutions to these challenges. Further, it identifies the available resources and open areas of research to aid those interested in Arabic IR research.

The remainder of this introductory section presents general interesting aspects of Arabic and outlines the content of subsequent sections in the survey.
1.1 The Arabic Language

Arabic is the most widely spoken Semitic language with an estimated 400 million speakers. Arabic shares many commonalities with other Semitic languages. These commonalities pertain to morphology, vocabulary, word order (subject-verb-object and verb-subject-object), use of short and long vowels, etc. For example, Arabic and Hebrew words are typically derived from roots that are composed of two, three, or four letters, with three letter (triliteral) roots being the most common. Words are constructed from roots by possibly inserting infixes, adding prefixes and suffixes, or doubling constants. Diacritics, which are often omitted in writing, help disambiguate words. Nouns can be singular, plural, or dual, and masculine or feminine.

Arabic has a broad sphere of influence which is mostly due to: a) religious reasons, where Arabic is the language of Islamic scholarship and that of the Muslim holy book, the Qur’an; and b) Arabic was the language of science and technology during the Middle Ages, with major Arabic universities in Spain, Africa, and the Middle East being learning hubs. Consequently, Arabic is part of school curricula in most majority non-Arab Muslim countries such as Iran and Pakistan. Arabic is also an official language in other countries such as Eritrea, Chad, and Somalia. Arabic had influence, mostly in terms of vocabulary, on many other languages such as Spanish, Turkish, Persian, Urdu, Swahili, and Hausa. Further, Arabic script is used for writing many languages such as Persian, Urdu, Kurdish, Pashto, and Dari.

The Arab population is generally a young population with an average age in the Arab World slightly less than 24\[1\] The Arabic language is ranked as the seventh top language on the web\[2\] The Arab Internet users constitute 3.3% of the Internet users worldwide. Although Arabic is ranked seventh among languages on the web, it is the fastest growing language on the web among all other languages (Figure 1.1). The number of Arab Internet users grew from 2.5 million in 2000 to 65 million in 2011. Internet penetration in the Arab World is estimated to

---

be 24%, which is lower than the global average of 32.7%. There are an estimated 45 million Arab Facebook users constituting roughly 5.6% of Facebook users globally. Though no exact estimates are available, Arabic online content is believed to constitute less than 1.5% of the global content. The relative size of Arabic forum content is disproportionately larger compared to the English forum content. English forum content is often considered of lower quality. However, such content is often of high quality in Arabic.

Modern Standard Arabic (MSA) is the lingua franca for the so-called Arab world, which includes northern Africa, the Arabian Peninsula, and Mesopotamia. Figure 1.2 shows a sample document written in MSA, which is an article from the Aljazeera.net news website. The article is written in MSA and would generally be understood by most Arabic speakers. However, Arabic speakers generally use dramatically different languages (or dialects) in daily interactions. There are six dominant dialects, which are Egyptian (85+ million speakers), Maghrebi (75+ million), Levantine (35+ million), Iraqi (25 million), Gulf (25+
1.1. The Arabic Language

Figure 1.2: Sample Arabic document from Aljazeera news website

Aside from those, there are tens of different Arabic dialects along with variations within the dialects. Due to the spread of social media, users are increasingly using Arabic dialects online. These dialects may differ in vocabulary, morphology, and spelling from MSA and most do not have standard spellings.

The fast growth of Arabic content on the web and the large variations between MSA and different dialects make it essential to develop effective IR systems. Figure 1.1 shows the number of users for some of the languages and their growth trends. In this survey, the efforts exerted for developing these systems are explored for different IR applications.

1.2 The Remainder of the Survey

The subsequent sections cover the following topics:

Section 2 entitled “Arabic Features Affecting Retrieval”, presents key aspects of Arabic that affect retrieval. These key aspects include Arabic orthography and morphology; the use of MSA vs. dialects; the differences between formal and informal text; the use of non-standard textual representations; and Arabic properties affecting the retrieval of content in different modalities, namely print and speech.

Section 3 entitled “Arabic Preprocessing and Indexing”, presents the core preprocessing steps that are required to prepare Arabic text for effective IR. The preprocessing steps including handling different encodings of Arabic, orthography, morphology, lexical and spelling variations, and stopwords. It also introduces effective index terms for Arabic.

Section 4 entitled “Arabic IR in Shared-Task Evaluations”, explores the presence of the Arabic language in different IR evaluation campaigns such as TREC, TDT, BOLT, and CLEF. It also presents the different IR tasks at the campaigns, namely ad hoc retrieval, filtering, cross-language retrieval, topic detection and tracking, and question answering.

Section 5 entitled “Domain-specific IR”, surveys work on different IR applications. These applications include cross-language IR, document image retrieval, general web search, social search, question answering, image retrieval, and speech search. The section addresses some of the challenges associated with different applications and some of the solutions that are reported in the literature.

Section 6 entitled “Open Research Areas in Arabic IR”, explores open areas of research that require more work. These areas include ad hoc IR, question answering, social search, and web search.

Section 7 concludes the survey.

Appendix A focuses on listing and providing links to Arabic resources that can be useful for IR such as test collections, stemmers, index tools, and translation tools.


Bibliography


Bibliography


Full text available at: http://dx.doi.org/10.1561/1500000031


Bibliography


Bibliography


[172] Raghavendra Udupa, K. Saravanan, A. Bakalov, and A. Bhole. 2009. „They are out there, if you know where to look”: Mining transliterations of OOV query terms for cross-language information retrieval. In the 2009 European Conference on Information Retrieval (ECIR), LNCS 5478, pp. 437-448.

Full text available at: http://dx.doi.org/10.1561/1500000031


