Turkish Information Retrieval: Past Changes Future

Fazli Can

Bilkent Information Retrieval Group,
Department of Computer Engineering,
Bilkent University, Bilkent, Ankara 06800, Turkey
canf@cs.bilkent.edu.tr

Abstract. One of the most exciting accomplishments of computer science in the lifetime of this generation is the World Wide Web. The Web is a global electronic publishing medium. Its size has been growing with an enormous speed for over a decade. Most of its content is objectionable, but it also contains a huge amount of valuable information. The Web adds a new dimension to the concept of information explosion and tries to solve the very same problem by information retrieval systems known as Web search engines. We briefly review the information explosion problem and information retrieval systems, convey the past and state of the art in Turkish information retrieval research, illustrate some recent developments, and propose some future actions in this research area in Turkey.

1 Introduction

The size of information has been growing with enormous speed. For example, it is estimated that in 2003 for each person on earth 800MB of information is produced. The majority of this information is boring such as supermarket scanner data. (Please also note that data, which is considered as boring by most people, can be interesting for data miners.) It is also estimated that 90% of currently produced information is in a digital form. It is expected that the most useful information will be in digital form within a decade [1].

Abundance of information has been a problem for a long time [2], [3]. Humans in their pursuit of truth, happiness, security, and prosperity have always chased the siblings "data, information, knowledge and wisdom." In the second half of the 20th century regarding the quantity of data, Donald E. Knuth writes "Sometimes we are confronted with more data than we can really use, and it may be wisest to forget and to destroy most of it..." Many of us do this successfully mostly by ignoring the available data or by conscious or unconscious selective attention. At the same time, we try to register and process as much information as possible, and produce a meaningful output, in the form of knowledge and finally wisdom. In this direction Knuth continues "... but at other times it is important to retain and organize the given facts in such a way that fast retrieval is possible" [4]. Herbert Simon indicates that the abundance of information creates poverty of attention: "... information ... consumes the attention of its recipients. Hence

T. Yakhno and E. Neuhold (Eds.): ADVIS 2006, LNCS 4243, pp. 13–22, 2006.

a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information resources that might consume it." [5].

The exponential growth of information is referred to as "information explosion" [3], [6], [7]. The abundance of information and the abundance of options provided by it create excessive stress on individuals in the form of information and decision overload [7]. Information retrieval systems, and more recently Web search engines, come to the rescue: these systems stretch our limits by storing and organizing information, and finally retrieving and prioritizing (ranking) relevant information when it is needed.

The goal of this paper is to review the information explosion problem and information retrieval process in general, convey the state of the art in Turkish information retrieval and some recent developments in that area, and propose some pointers for future actions in Turkey.

2 Information Explosion and Information Retrieval Systems

Information explosion is a long-term phenomenon. For example, in 1945, Dr. Vannevar Bush in his frequently cited classic article "As we may think" indicated that society was creating information much faster than it could use. Bush was then headed six thousand scientists in the application of science to warfare in the US [2]. In his article, Bush imagined a mechanized private file and library called "Memex" for personal information management. Memex was imagined as a device in which an individual stores all his books, records, communications, photographs, memos, etc. that can be consulted with "exceeding speed and flexibility." It can be seen as a forerunner of the present day information retrieval systems.

Fig. 1. An example of personal information explosion and brute force solution to problem (boxes on the left mostly contain personal documents)

Today many people experience information explosion first in their personal lives: as individuals we have to deal with many documents related to our family members and ourselves. We have to keep and organize these documents for

possible future needs, for example, to prove a payment. Figure 1 illustrates the "information explosion" problem that I experienced and my brute force solution to that problem. In years, we accumulate a good amount of paper documents: tax forms and related papers, insurance policies, health related documents, receipts and statements, cancelled checks, etc. In years, this accumulation can reach to an unnecessarily huge size. The left picture of Figure 1 shows the physical evidence of the problem in my case. In this picture, the boxes mostly contain aforementioned documents. As a solution to this, I went over these boxes, filtered the necessary items - a small amount- and shredded the rest as shown in the right picture. In the second picture, next to the shredder, only one of the many bags is shown and in my case, this process took several days. Paper shredders are a kind of Occam's razor [8]: a device that simplifies our lives by safely eliminating unnecessary documents.

In our daily lives, in addition to paper we have huge amount of digital information: digital pictures and movies, emails, papers in various electronic formats, news articles, etc. Handling them effectively and efficiently is not easy. For keeping our personal data in order, we start to see the emergence of a new technology, a new kind of information retrieval system, called personal information management systems. In the future, such systems may even provide a total recall of our lifetime experiences [9].

The size of the Web provides another example for the "information explosion" phenomenon. Regarding the overall size of the Web, or on the coverage of the Web by search engines, we see continuously increasing numbers. Finding the actual Web size or its coverage by Web search engines is difficult and beyond the scope of this article (a good resource about this is searchenginewatch.com).

Information retrieval systems aim to locate documents that would satisfy a user's information needs. Here we limit our discussion to retrieval from natural language text. Users of such systems usually specify their information needs using a few words. The information retrieval research field was emerged in 1950s as a part of computer science and information science. Calvin Moores, a pioneer of information science, coined the term in 1951; Gerard Salton, a computer scientist, is known as the father of modern information retrieval [6].

Since document collections are very large, IR systems perform retrievals on document representatives. Various models exist for document representation one of which is the vector space model [10]. In the vector space model, a document collection can be represented by an imaginary document by term matrix. In this matrix, each row represents a single document as a collection of terms. Each row element is called an index term. Usually this matrix is stored as an inverted index structure that contains a posting list for each term used in the documents [8]. Each posting list contains a list of documents containing the corresponding term.

During indexing, terms are assigned weights (importance) according to their occurrence patterns in individual documents and collection. The importance of a term in a document is usually proportional to its number of occurrences in that particular document (indicated by *tf* - term frequency). Term importance

is inversely proportional to its collection frequency (indicated by *idf* - inverse document frequency), that is, a term that appears in several documents are assigned a lower weight since such terms are not good at discriminating documents from each other during the retrieval process [8] [10]. By using query-document matching functions documents containing more query terms with higher weights are listed first. In ranking, Web search engines may also take advantage of the hypertext link structure available on the Web [11].

3 Research on Turkish Information Retrieval

In this section, we provide a short survey of the research done on Turkish information retrieval. The coverage may be incomplete; however, still a good representative of the published studies. On the Web there are many Turkish Web search engines/directories (after a simple search we were able to identify about thirty of them). Their quality and coverage vary. We keep them out of our concern since they conceal their retrieval techniques [12].

The first component of most IR related research is test collection. IR test collections consist of three parts: a set of documents, a set of user information requests or queries, and the set of relevant documents for each query. Standard test collections facilitate reproducibility of results and easy comparison among the performance of different retrieval techniques. The major concern of IR research is effectiveness. In information retrieval measuring effectiveness involves two concepts: precision and recall. Precision is proportion of retrieved documents that are relevant, and recall is proportion of relevant document retrieved. Other effectiveness measures used in IR are usually the derivatives of these two concepts [13].

The earliest published Turkish IR study, which is done by Köksal, uses 570 documents (title, keywords, section titles, and abstract) on computer science with twelve queries. It measures the effectiveness of various indexing and document-query matching approaches using recall precision graphs and uses a stop list of size 274 that includes frequent words of Turkish (such as "bir, ve") in order to not to use them in the retrieval process. For stemming purposes, Köksal uses the first five characters (5-prefix) of words. This selection is done after experimenting with various prefix sizes [14].

Solak and Can use a collection of 533 news articles and seventy-one queries. For stemming, a morphological parser has been used and the study uses several query-document matching functions. The study shows effectiveness improvement with stemming with respect to no stemming. The reported experiments employ seven different term weighting approaches [15].

Sever and Bitirim describe the implementation of a system based on 2468 law documents and fifteen queries. First, they demonstrate the superior performance of a new stemmer with respect to two earlier stemmers (one of them is the Solak-Can stemmer mentioned above). Then they show that their inflectional and derivational stemmer provides 25% precision improvement with respect to no stemming [16].

Table 1. Turkish IR test collections

Researcher(s), Year	Contents	No. of Documents	No. of Queries
Köksal , 1981	Computer Science	570	12
Solak, Can, 1994	Newspaper articles	533	71
Sever, Bitirim, 2003	Law documents	2468	15
Pembe, Say, 2004	Various topics from Web	615	5

Pembe and Say study the Turkish information retrieval problem by using knowledge of the morphological, lexico-semantical and syntactic levels of Turkish. They consider the effects of stemming with some query enrichment (expansion) techniques. In their experiments, they use 615 Turkish documents about different topics from the Web and five long natural language queries. They use seven different indexing and retrieval combinations and measure their performance effects [17]. For easy reference, Table 1 provides the characteristics of the Turkish IR test collections mentioned above.

If we consider the research done in information retrieval for the English language, we see two distinct periods. These are the pre-TREC and the TREC periods. TREC, Text Retrieval Conference, is co-sponsored by the National Institute of Standards and Technology (NIST), the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA/ITO), and the Department of Defense Advanced Research and Development Activity (ARDA) of the United States. The first TREC conference was held in 1992. TREC workshop series aims: "a) to encourage research in information retrieval based on large test collections; b) to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas; c) to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and d) to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems" [18]. The TREC conference had a remarkable effect on the quality and quantity of the IR research.

Before TREC the following (English) test collections were commonly used (name, size in number of documents, number of queries): (ADI, 82, 35), (CACM, 3200, 64), (INSPEC, 12684, 84), (NPL, 11429, 100), and (TIME, 423, 83) [13] [19]. Compared with Turkish collections these are mostly larger and involve significantly more number of queries. The TREC collection sizes change depending on factors such as needs, application and availability of data. Information for "some" of the TREC collections used for ad hoc information retrieval based on a few query words is as follows: (WSJ-Wall Street Journal 1987-1989: 98,732 documents), (FT-Financial Times 1991-1994; 210,158), (FR-Federal Register 1994: 55,630). These TREC collections are used with 50 queries and their sizes respectively are 267, 564, 395MB [13]. A new TREC collection is the GOV2. It contains Web data crawled from the .gov domain. It is 426GB in size and contains approximately twenty-five million documents [18]. Different from ad hoc searches

the use of this collection involves meta data in the query (topic) statements. Experimental results obtained by using large test collections like those of TREC would be easier to generalize to real world cases.

4 Turkish Information Retrieval Changes: New Developments

In Turkish information retrieval, we have a new research project undertaken by the Bilkent Information Retrieval Group [20]. It aims to investigate various aspects of Turkish information retrieval in large-scale dynamic environments. One of the goals of the group is to develop a Turkish news portal that would provide information retrieval, information filtering, new event detection and tracking, and output clustering and summarization services [21].

The first step of this effort is creating a TREC-like large standard test collection for Turkish information retrieval and measuring the performance of several retrieval techniques that involve various stemming and query-document matching functions. The test bed, which has been created for this study, contains 95.5 million words including numbers (1.3 million distinct words and 0.1 million distinct numbers), 408,305 documents and has a size of about 800MB. All documents come from the Turkish newspaper *Milliyet* (www.milliyet.com.tr). The collection contains news articles including columns of five complete years, 2001 to 2005. We also have seventy-two Web-like ad hoc queries created by more than thirty participants by spending more than total of three days of query evaluation time. The relevant documents of the queries are determined by using a TREC-like pooling approach [20].

In the experiments we use eight query-document matching functions (MF1 to MF8) based on the vector space model. The first one, MF1, is the well-known cosine function. MF1 involves no *idf* component just computes the cosine of the angle between query and document vectors within a multi dimensional space. The matching functions, MF2 to MF7, are highly recommended in [19]. Finally, MF8 [8] reflects the *idf* effects of collection changes to query term weights and requires no change in document term weights as the collection size changes and therefore especially suitable for dynamic environments.

In the experiments, we use four stemming options: no stemming (NS), first-n characters (Fn) of each word, the successor variety (SV) method [22], and a lemmatizer-based stemmer supported by a morphological analyzer [23] for obtaining more accurate stems [24]. The Successor Variety, SV, algorithm determines the root of a word according to the number of distinct succeeding letters for each prefix of the word in a large corpus. The expectation is that the stem of a word would be the prefix at which the maximum successor variety, i.e., the distinct number of successor letters, is observed. On the other hand, a lemmatizer identifies the "lemma" of a word, i.e., its base form in dictionary. For a given word, a lemmatizer can provide more than one alternative. In such cases, we choose the alternative whose length is closest to the average lemma length (6.58 characters) of word types. If there are multiple candidates we choose the one whose corresponding

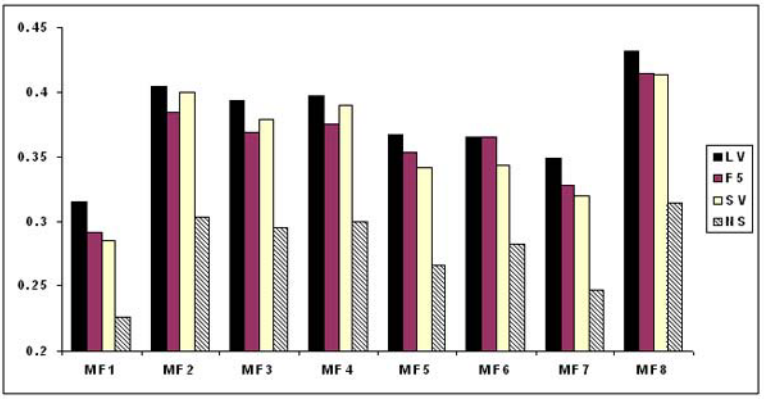

Fig. 2. Retrieval effectiveness (bpref) of matching functions MF1-MF8

part of speech, POS, information is most frequent in Turkish. It is experimentally shown that this approach is more than 90% accurate [24]. In choosing lemmas, we also use the length of 5 characters instead of 6.58 (since retrieval with F5 gives good results). This way, we have two lemmatizer-based stemmer versions: LM5 and LM6. For miss spelled and foreign words, which cannot be analyzed by the lemmatizer (about 40% of all distinct words), in an additional **L**M5 version we use the S**V** method for such words, this version is referred to as LV.

The queries are created according to the TREC ad hoc query tradition using binary judgments. The relevant documents are identified by taking the union of the top 100 documents of the twenty-four possible retrieval combinations, "runs," of the eight matching functions and the stemmers NS, F6, and SV. In our experiments for measuring effectiveness, a relatively new measure, bpref, has been used to prevent any possible bias effect on the systems not involved in query pool construction [25]. The bpref is designed especially to handle cases like this and can have a value between 0 and 1, where 1 is the best possible value which indicates that all relevant documents appear at the beginning of the ranked query results. For the final analysis, we have LV, F5, SV and NS. The other cases are not included in the final evaluation process due to their poor or similar performances to these stemmers, NS is our baseline case. Figure 2 shows the results and illustrates that NS (no stemming) is much worse than the others. The most effective one is LV. The SV method and the simple prefix method F5 are also effective, but not as good as LV. The comparisons involve statistical tests as reported in [20].

The experiments show that truncating words at a prefix length of 5 provides an effective retrieval environment in Turkish. However, a lemmatizer-based stemmer provides better effectiveness over a variety of matching functions. Our TREC-sized test collection for Turkish, which we plan to share with other researchers, is one of the main contributions of this project. Currently our group is working on experiments such as query length effects on Turkish information retrieval and the scalability issues.

In addition to our experimental evaluation, we also have an operational system called BIRnews (Bilkent Information Retrieval -Group- News) which is based on our experimental findings . This is the first step towards the multi functional news portal that we plan to implement. By using this system, users can search our *Milliyet* news archive. BIRnews is available on the Web at the following address: `http://bilkent.edu.tr/birnews`. The current advanced users's interface of the system allows users to experiment retrieval with various stemmer and matching function combinations.

5 Conclusions

Information retrieval systems can be used to control how (which, why, when, and where) things are remembered. In other words, these systems can have a bias in terms of what they retrieve and present to their users. They can affect or even control how we perceive, think, and decide. Web search engines can do this simply for advertisement or their bias can be due to their Web crawling and indexing decisions [26]. In addition to these, Web sites can try to embed their own bias to the retrieval process. This is done by techniques known as search engine optimization. Such techniques try to make some Web pages more accessible, i.e., ranked higher in search results.

It may be an old cliché, but it is true that "information is a valuable commodity." Effective information retrieval systems provide better communication between information resources and receivers. Such systems can have a significant impact on improving society and making it more prosperous and better educated. Furthermore, in several applications (one good example is "national security") we would need to have information retrieval systems that could retrieve data not only effectively and efficiently, but also objectively without any bias. Most IR research findings could be language independent and therefore universal. However, when we look at the research done on Turkish information retrieval, although we have some efforts, this research area still looks like an uncharted land. We need to explore and claim this territory. This can be done by

- promoting intra- and inter-institution collaboration among researchers,
- encouraging research and development for applications ranging from personal information management to national digital library development,
- generating communication among different groups by creating an open forum for the exchange of research and development ideas.

In Turkey, we need a TREC-like initiative to promote, and support these important actions. This should be done in an organized manner. For this purpose, governmental institutions and non-governmental organizations can provide support and resources.

Acknowledgements. I am grateful to late Prof. Esen A. Özkarahan; my friend, teacher, Ph.D. advisor, mentor, and colleague; who traveled with me and introduced me to the field of information retrieval. I would like to thank Ismail Sengör

Altingövde for his valuable comments on an earlier version of this paper. This work is partially supported by the Scientific and Technical Research Council of Turkey (TÜBITAK) under the grant number 106E014. Any opinions, findings and conclusions or recommendations expressed in this article belong to the author and do not necessarily reflect those of the sponsor.

References

1. Varian, H. R.: Universal Access to Information. Com. of the ACM **48** (10) (2005) 65-66
2. Bush, V.: As We may Think. The Atlantic Monthly **176** (1) (1945) 101-108
3. de Solla Price, D.: Little Science, Big Science... and Beyond. Columbia University Press, New York, 1986 (originally published in 1963)
4. Knuth, D. E.: The Art of Computer Programming, volume 3: Sorting and Searching. Addison-Wesley, Reading, MA (1973)
5. Stefik, M.: The Internet Edge. MIT Press, Cambridge, MA, 1999
6. Saracevic, T.: Information Science. Journal of the American Society for Information Science **50** (12) (1999)1051-1063
7. Toffler, A.: Future Shock. Bantam Books, New York , 1990 (originally published: 1970)
8. Witten, I. H., Moffat, A., Bell T. C.: Managing Gigabytes Compressing and Indexing Documents and Images, 2nd edition. Morgan Kaufmann Publishers, San Francisco (1999)
9. Gemmell, J., Bell, G., Lueder, R.: MyLifeBits: a Personal Database for Everything. Com. of the ACM **49** (1) (2006) 89-95
10. Salton, G..: Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer. Addison Wesley, Reading, MA (1989)
11. Brin, S., Page, L.: The anatomy of a large scale hypertextual Web search engine. Computer Networks and ISDN Systems, **30** (1-7), (1998) 107-117
12. Bitirim, Y., Tonta, Y., Sever, H.: Information Retrieval Effectiveness of Turkish Search Engines. In: Yakhno, T. (ed.): Advances in Information Systems. Lecture Notes in Computer Science, Vol. 2457. Springer-Verlag, Berlin, Heidelberg New York (2002) 93-103)
13. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading, MA (1999)
14. Köksal, A.: Tümüyle Özdevimli Deneysel Bir Belge Dizinleme ve Erisim Dizgesi: TÜRDER. In the Proceedings of 3. Ulusal Bilisim Kurultayi, Ankara, Turkey. (1981) 37-44
15. Solak, A., Can, F.: Effects of Stemming on Turkish Text Retrieval. Int. Symposium on Computer and Information Sciences (ISCIS), (1994) 49-56
16. Sever H., Bitirim, Y.: FindStem: Analysis and Evaluation of Stemming algorithms for Turkish. In: String Processing and Information Retrieval. Lecture Notes in Computer Science, Vol. 2857. Springer-Verlag, Berlin, Heidelberg New York (2003) 238-251
17. Pembe, F. C., Say, A. C. C.: A Linguistically Motivated Information Retrieval System for Turkish. In: Aykanat, C., Dayar, T., Korpeoglu, I. (eds.): Computer and Information Sciences. Lecture Notes in Computer Science, Vol. 3280. Springer-Verlag, Berlin, Heidelberg New York (2004) 741-750

18. Voorhees, E.: Overview of TREC 2004. `http://trec.nist.gov` (accessed on June 16, 2006)
19. Salton, G., Buckley, C:. Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management **24** (1988) 513-523.
20. Can, F, Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., Vursavas, O. M: First Large Scale Information Retrieval Experiments on Turkish Texts. (Poster paper) In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, (2006), to appear
21. Radev, D., Otterbacher J., Winkel, A., Blair-Goldensohn, S.: NewsInEssence: Summarizing Online News Topics. Com. of the ACM **48** (10) (2005) 95-98
22. Hafer, M. A., Weiss, S. F.: Word Segmentation by Letter Successor Varieties. Infor. Stor. Retr. **10** (1974) 371-385
23. Oflazer, K.: Two-level Description of Turkish Morphology. Literary and Linguistic Computing, **9** (2) (1994) 137-148
24. Altintas, K., Can, F., Patton, J. M.: Language Change Quantification Using Time-Separated Parallel Translations. Literary and Linguistic Computing (accepted)
25. Buckley, C., Voorhees, E. M.: Retrieval Evaluation with Incomplete Information. In Proceedings of the 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval. (2004) 25-32
26. Mowshowitz A, Kawaguchi A.: Assessing Bias in Search Engines. Information Processing and Management **38** (1) (2002) 141-156