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