Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems ••• (••••) •••–•••

FUZZY sets and systems

www.elsevier.com/locate/fss

Aggregation operators in Information Retrieval

Stefania Marrara, Gabriella Pasi, Marco Viviani

Università degli Studi di Milano-Bicocca / DISCo, Viale Sarca, 336, 20126 Milano, Edificio U14, Italy

Received 26 February 2016; received in revised form 23 December 2016; accepted 26 December 2016

Abstract

Information Retrieval is a complex task, which deals with both the subjectivity related to the user's needs and the uncertainty and vagueness that characterize the retrieval process. For this reason, deciding to which extent a document is relevant to a user's needs is not easy, and it strongly depends on several dimensions such as topicality, novelty, user's context, and so on. One of the most straightforward ways to interpret this activity is as a Multi-Criteria Decision Making (MCDM) problem, in which the choice of appropriate aggregation operators can play an important role in various tasks related to, or characterizing the IR process. This article aims to provide a presentation of the main approaches that in the literature have made use of aggregation operators in Information Retrieval.

Keywords: Aggregation operators; Information Retrieval; Indexing; Query languages; Multidimensional relevance assessment; Metasearch; OWA operators; Copulas; Choquet integrals

1. Introduction

In last years the increasing expansion of the Web, with the consequent availability of a huge amount of on-line content, has motivated a big deal of research aimed at the definition of systems that may help users in locating information relevant to their interests and preferences. In this scenario, search engines (also called Information Retrieval Systems) play an important role. In particular, the goal of a Web search engine is to retrieve Web pages relevant to information needs that the user expresses in a query by means of a few keywords.

Information Retrieval is a subjective task, and as such various phases of a retrieval process are affected by uncertainty and/or vagueness. As is well known, when using a search engine the query formulation process is subjective, and the user may be uncertain on which keywords to select to properly express her/his needs. In fact, the user has often only a vague idea of what s/he is looking for, as extensively pointed out in the literature. It is today widely recognized that Information Retrieval is an interactive process. A manifestation of the interactive nature of search is that users often engage in query sessions where they specify a sequence of refined queries to the aim of improving the specification of their needs. Due to the complexity of the notion of relevance of an information item to some specific

E-mail addresses: stefania.marrara@disco.unimi.it (S. Marrara), pasi@disco.unimi.it (G. Pasi), marco.viviani@disco.unimi.it (M. Viviani).

http://dx.doi.org/10.1016/j.fss.2016.12.018
0165-0114/

needs, a search engine can only estimate it; moreover, a search engine makes such an estimate on the basis of several relevance dimensions. The first search engines (including the first Web search engines) estimated relevance only based on topicality, i.e., the topical affinity of a query to the content of a document. While topicality still constitutes the core relevance dimension, other dimensions have been deemed important to the assessment of the relevance of a document to a query. For example, in Web search the popularity of a Web page has been first proposed by Google as a measure of the quality of the page, independently of its content [1]. Additional dimensions that can be taken into account to assess the relevance of a document to a query include: the recency of a document, its novelty, the location of the user in case of query formulated on mobile devices, etc. More recently, personalized search approaches are aimed at producing search results that are tailored to the user's context and preferences, formally represented in a user model (user profile). The above mentioned facets of the notion of relevance give an idea of the fact that the effectiveness of a search engine is crucially related to its capability of accounting for the contextuality, the vagueness, and the uncertainty of the retrieval process.

In this article, we focus on approaches to IR that have addressed the issue of defining flexible IR systems by employing aggregation operators in various phases of the retrieval process. In particular, these approaches rely on the interpretation of IR as a Multi-Criteria Decision Making (MCDM) problem, from various perspectives. The first, more straightforward perspective is to interpret the overall IR process as an MCDM process aimed at selecting the best alternatives (documents) based on the assessment of the performance of multiple criteria (i.e., the keywords specified in a user's query). Another and strongly related perspective is to see the assessment of the overall relevance estimate of a document (still an alternative) to a query as the process of evaluating the performance of several relevance dimensions (e.g., topicality, novelty, recency, etc.), which in this case represent the considered criteria.

As it will be seen in this article, other processes that characterize an Information Retrieval System can be interpreted as MCDM problems, thus requiring the application of appropriate aggregation operators: this is the case of the indexing process when applied to structured documents.

Metasearch constitutes another interesting task that can be seen as an instance of a Multi-Expert Decision Making problem, also strongly relying on the appropriate choice of an aggregation operator. By this task, a user query is separately evaluated by different search engines, each one providing its relevance assessment of the considered documents. Metasearch aims to merge the relevance assessments made by the various search engines (experts) to the aim of providing a unique, consensual ranked list of documents to the user in response to her/his query.

A quite interesting aspect implied by the above interpretations of the IR process (or of some of its components) is that the choice of distinct aggregation operators can produce different results. In other words, the semantics of aggregation implies an interpretation of the affected process.

Despite the potential impact of aggregation in IR, this aspect has not received the proper attention in the literature. Only recently some approaches have appeared demonstrating the importance of this problem, and its impact on an IRS. This article aims to review the main contributions that in the literature have made use of aggregation operators in Information Retrieval.

The article is organized as follows: after a brief presentation of Information Retrieval and of the main characteristics of an Information Retrieval System in Section 2, the relationship between Information Retrieval and aggregation operators is outlined in Section 3. Section 4 describes the main approaches employing aggregation-based techniques in document indexing. The use of aggregation operators in the definition of query languages and in query evaluation is presented in Section 5. Section 6 and Section 7 discuss respectively the use of aggregation operators in assessing multidimensional relevance and in metasearch. Finally, Section 8 draws out the conclusions.

2. Information Retrieval

Information Retrieval (IR) addresses the issues of representing, storing, and accessing huge document collections. The aim of an Information Retrieval System (IRS) (or search engine) is to evaluate a user's query and to retrieve the documents that it estimates *relevant* to that query. In the following, we describe the main components of an IR system. First, an IRS provides a formal representation of both users' queries and documents. Such a representation relies on a formal model; in the literature several models have been proposed, among which the *Boolean model* [2], the *Vector Space Model* (VSM) [3], and several *probabilistic models* [4], just to name the most popular. In Fig. 1 the main components of an IR system are sketched:

Please cite this article in press as: S. Marrara et al., Aggregation operators in Information Retrieval, Fuzzy Sets Syst. (2016), http://dx.doi.org/10.1016/j.fss.2016.12.018