Fuzzy Rough Set Based Technique for User Specific Information Retrieval: A Case Study on Wikipedia Data

Niladri Chatterjee, IIT Delhi, Delhi, India
Nidhika Yadav, IIT Delhi, Delhi, India

ABSTRACT

Information retrieval is widely used due to extremely large volume of text and image data available on the web and consequently, efficient retrieval is required. Text information retrieval is a branch of information retrieval which deals with text documents. Another key factor is the concern for a retrieval engine, often referred to as user-specific information retrieval, which works according to a specific user. This article performs a preliminary investigation of the proposed fuzzy rough sets-based model for user-specific text information retrieval. The model improves on the computational time required to compute the approximations compared to classical fuzzy rough set model by using Wikipedia as the information source. The technique also improves on the accuracy of clustering obtained for user specified classes.

KEYWORDS

Fuzzy Rough Sets, Fuzzy Sets, Information Retrieval, Lower Approximation, Natural Language Processing, Rough Sets, Soft Computing, Text Processing, Upper Approximation

INTRODUCTION

Information retrieval (IR) (Luhn, 1955) concerns with extraction and ranking of required data from a corpus which is most relevant to a particular user specified query. In a broad sense the corpus can be a collection of text documents under consideration, textually tagged image files, the intranet/internet documents or images. When the data under consideration is in the form of text the process of extraction for text files which are important to a query is called text information retrieval. Text information retrieval is crucial in these days of information abundance. This field has gained immense popularity keeping in view the availability of text in the form of documents, blogs, tweets, product reviews, news articles, official documents in an organization and websites to mention few. Another equally important and growing field is IR of images based on textual queries. This field is growing in both research and industry in various applications and is popularly referred to as content-based information retrieval (Datta et al., 2017). Though a relatively new area as compared to its text counterpart, it is of prime concern keeping in view the cheap availability of cameras in phones, laptops and other digital devices.
devices. This lead to rise of image data with textual meta data present in social media applications such as, Instagram, Twitter, Facebook etc.

A large number of developments has been made in the field if IR in last fifty years. Luhn (1955) proposed one of the first IR systems wherein exact match were computed between the query and document. Since then various refined and advanced IR techniques have been proposed and in use in modern day IR systems. Salton and Bukley (1988) in their seminal work described various term weighting techniques that could be used for an efficient IR system. The query and document were represented as vectors and similarity between the query and document was computed as dot product between them. IR has matured now to a more sophisticated system using advanced mathematical and Artificial Intelligence techniques, such as Google PageRank algorithm, Deep Learning concepts.

In an IR system it is the user who plays the key role. Given the same query by two different users an unbiased IR engine shall produce the same results. It is worth noting that two different users imply different interests and choices. As a consequence, it is paramount to take into consideration the personal choices, the interests and habits of a particular user. Hence, there is a necessity to develop a client-side filtering system. This system filters the information retrieved by an IR engine, e.g. Google, Bing, Yahoo to name a few. Ones the data has been retrieved by IR engine filtering is performed at client side. The present work is concerned with the process of filtering the text data Information Retrieval which is specific to a particular user. Text IR henceforth will be referred to as IR in this paper. User-specific IR can improve on the search results significantly since not all information retrieved by a search engine is relevant to all users. This can be illustrated with the help of an example. Consider two different users: one who is into fishing business and other who is a musician. Both these users enter the keyword “bass”. We note that both refer to different contexts even though the keyword is same. The first one refers to the “bass fish” while the second one refers to the instrument “bass.” Once a user specific IR system is built at client side only the information that is relevant to this user shall be retrieved. Hence it is of prime importance to learn users’ opinion for an efficient IR system.

Another aspect that the current research focus on is that the queries used for retrieval are typically of few words in length which makes it difficult to retrieve relevant documents with precision. The query term may not be as such present in the document, though may be present in a document in the form of synonyms or some other lexical relations, such as hypernym, hyponym. Rather, a document can have an association to query based on information extracted from external database, such as WordNet, Wikipedia among others. The present work is based on this concept. WordNet has been used for extracting the relations present between query and document. The present work extends logic to use Wikipedia as a source of information for computations of similarity between query and document. Wikipedia is used as a rich source of information, as the external knowledge base. The reason behind choosing Wikipedia as an information source is that Wikipedia has plethora of information that can be utilized to compute relations between words of a document and a query.

Fuzzy rough set-based techniques have been proposed in literature for user specific IR (Singh and Dey, 2005, Srinivasan, 2001). The drawback of the classical approach of Fuzzy Rough set-based techniques is the huge temporal requirements in computing all the semantic similarities between all the words in WordNet over all the words present in a document. This was the motivation for the use of proposed approach using a complementary definition of fuzzy rough sets (Jensen and Shen, 2004), which has been used for feature selection.

In this paper user relevance feedback is extracted and henceforth used to rank documents pertaining to a query. The user feedback depends on the document, the user and hence associated with some amount of uncertainty (Alhabashneh et al., 2017). This research focus on handling these uncertainties using fuzzy rough set-based IR. Two kinds of relevance feedbacks are available in literature: explicit and implicit. The explicit feedback asks users to provide the feedback specifically, while the implicit feedback computes the relevance of a particular document or website based on various algorithm specific factors like reading time, response time, number of mouse clicks, mouse movements to mention few. various studies have been made in this respect in literature. The interests
Evaluation of Web Accessibility: A Combined Method
www.igi-global.com/chapter/evaluation-web-accessibility/70722?camid=4v1a