Applied Mechanics and Materials Vols. 519-520 (2014) pp 857-862
Online available since 2014/Feb/06 at www.scientific.net

doi:10.4028/www.scientific.net/AMM.519-520.857

Term weighting: A Multi-View Fuzzy ontology based approach

Zeinab E. Al-Arab[a], Ahmed M. Gadallah[b] and Hesham M. Hefny[c]

[a]eng.zeinabezz@gmail.com, [b]ahmgad10@yahoo.com, [c] hehefny@ieee.org

Keywords: Term weighting; Fuzzy ontology; Document annotation

Abstract. The paper proposes a term weighting algorithm for research papers. It weighs a research paper's annotated keywords according to a certain view. It uses a predefined multi-view fuzzy ontology and a stemmer NLP tool. The proposed algorithm is tested and results are compared with another ontology-based term weighting algorithm. The tests show that it enhances the resulted weights accuracy and decreases the execution time.

Introduction

An information retrieval system (IR) consists of a set of documents, a user query, a retrieval engine and a ranking module. It stores and indexes documents such that when users express their information needs in a query, the retrieval engine retrieves a set of relevant documents associating a score to each one. The higher the score is, the greater the document relevance. Then the ranking module ranks them and displays them to the user [1]. In order to index a set of documents, an annotation algorithm should be used.

Annotation is a process of adding metadata to a certain document in order to describe it. This metadata may include a string of weighed keywords, authors' names, the publishing conference or journal, date of publishing …etc. The weight of each keyword reflects to what extend it represents that document. The process of weighting a document's keywords is known as Term Weighting (TW).

This paper presents a domain-specific Term Weighting approach based on a multi-view fuzzy ontology. The paper is organized as follows: the next section presents term weighting technique. A survey about fuzzy ontology is conducted in section 3. Section 4 presents some related works. The proposed term weighting algorithm is presented in section 5. The proposed algorithm is discussed and concluded in Section 6.

Term weighting (TW)

Term weighting (TW) is to calculate a weight for each term representing a certain document. This weight reflects to what extent this term represent that document. Due to its importance, term weighting is used in many fields such as document clustering, information retrieval (IR), and many more. Regarding IR, it enhances the recall and the precision measure. Also it enhances the rank of the retrieved documents[2].

There are several algorithms implementing the term weighting concept. These algorithms are applied either to a domain specific or to a general one. Almost all general term weighting algorithms are statistical algorithms. One of the most popular term weighting algorithms for the general Term Weighting approach is the Term Frequency-Inverse Document Frequency, TF-IDF. Domain specific term weighting algorithms use domain ontology to expand a certain document keywords with their synonyms to increase their weights accuracy. These weights are calculated using some statistical formulas.

Fuzzy Ontology

Fuzzy ontology reprisernts uncertain information commonly found in many application domains in a human understandable, machine readable format[3]. It is used as a standard knowledge representation for the semantic web [4]. Although fuzzy ontology is used in many applications such as semantic web,

multiagent systems, information retrieval, and many more, no standard components for fuzzy ontology are found in the literature [5]. Researchers define fuzzy ontology components according to their used application and domain. Some of such definitions are as follows:

- Fuzzy ontology is a quadruple(C, R, P, I), where C is a set of fuzzy concepts, R is a set of binary relations, P is a set of fuzzy properties of concepts, and I is a set of individuals [6].
- Multi-View Fuzzy Related Ontologies, MVFRO, is a couple of (f-os, f-oI), where f-os is a fuzzy ontology structure, while f-oI is the fuzzy ontology individual of concepts and relationships associated with the fuzzy ontology structure. F-os is a quintuple (C, C_R, P, T, A). C is a set of fuzzy concepts. C_R is a fuzzy relation between concepts. It can have more than one value; each represents a certain view. P is a set of concept properties. T is a set of terms that express the concept c. A is a set of axioms [7].

Related work

Two annotation techniques based on crisp ontology are proposed in [8]. The first annotation technique uses an NLP annotation algorithm to annotate a certain document with a string of keywords. Then, these keywords are weighed using an adapted TF-IDF algorithm. This adapted algorithm is the frequency of the occurrence of each semantic entity in the ontology or any of its associate keywords within a document. Such an algorithm takes pronoun into account. The second technique uses a contextual semantic information based algorithm to annotate a certain document with a string of keywords. Then, these keywords are weighed using a fusion weighting algorithm.

An annotation system performing a clustering process based on a concept weight supported by crisp domain ontology is proposed in [9]. The system is divided into three major modules; document preprocessing, calculating a concept weight based on ontology, and clustering documents with the concept-based. The weighting module is calculated by the Eq.1 [9]:

$$W = Len \times Frequency \times Correlation\ Coefficient + Probability\ of\ concept \quad (1)$$

Where, W is the weight of a certain keyword. Len is the length of that keyword. Frequency is times which the words appear. Probability is based on the probability of the concept in the document.

A new weighting method based on statistical estimation of a word importance for a particular categorization problem is proposed in [10]. This weighting also has the benefit that it makes feature selection implicit since useless features for the categorization problem considered get a very small weight.

The proposed term weighting algorithm

The proposed algorithm is a semantic-based term weighting algorithm. It calculates a weight for each annotated keyword in a certain paper according to a specific view. This weight reflects to what extend a certain keyword represents a specific paper according to a specific view. It uses a multi-view fuzzy ontology and some Natural Language Processing (NLP) tools. The algorithm aims to enhance the resulted weights accuracy, and decrease the execution time. Enhancing the resulted weights accuracy can be achieved through:

- Using a multi-view fuzzy ontology for expanding each annotated keyword in the keyword zone in the required view,
- Arranging the paper-expanded keyword list in a descending order according to each keyword n-grams and the number of terms in each keyword,
- Replacing each pronoun with its referred noun, instead of removing it as stop word,

- Working only on paragraphs' main sentences as they reflect the paragraph main idea, instead of working on the whole paper.
- For entitled sections, they are annotated with their titles not with keywords listed in the paper keyword zone.

Decreasing the execution time can be achieved through considering each paragraph's main sentence, instead of considering all document words.

1.1. The Proposed Term Weighting Algorithm Phases

The proposed algorithm phases are as follow:

1- Preprocessing phase

Firstly, the paper is divided into different weighted zones. Zone is one of the standard sections in any research paper, e.g., Title, Abstract, Introduction, Related Work, References, etc. The weight of each zone reflects the role of this zone in the paper, e.g., {(title, 1), (introduction, 0.5)……}.

Secondly, remove the reference zone from the paper. Then, extract the annotated keywords from the keyword zone associating each of them with a weight, w.

$$PKS= \{(k_{i0}, w_{ki0})\}, i=1….n,$$

Where PKS is the Paper Keyword Set; k_{i0} is a keyword in the keyword zone. n is the number of keywords in the keyword zone. W_{ki0} reflects to what extent k_{i0} represents this document. The value of $w_{ki0} \epsilon$]0, 1] is calculated through this algorithm.

2- Annotating paper Zones

Each zone the paper is annotated with the PKS and its expansion set. This phase is responsible for expanding each keyword in the PKS with all concepts and all terms related to it with a certain threshold in specific view using the predefined multi-view fuzzy ontology. Each keyword is stemmed and then expanded through the following steps:

- If it is represented in the fuzzy ontology as a concept, expand it with all its related concepts, and terms that can represent it with degree greater than or equals to a certain threshold in the considered view.
- If it is represented in the fuzzy ontology as a term, expand it with all its related terms, and concepts that it can represent in the document domain with a degree greater than or equals to a certain threshold in the considered view.

So that,

$$PKS= \{(k_{ij}, w_{kij})\}, i= 1…n, j=0…..m,$$

Where, k_{i0} is a keyword in the keyword zone.w_{ki0} reflects to what extent k_{i0} represents this paper, The value of $w_{ki0} \epsilon$] 0, 1] is calculated through this algorithm. n is the number of keywords in the keyword zone. k_{ij}, j=1…m, is an expanded keyword for the keyword k_{i0}. m is the number of the k_{i0}'s expanded keywords. w_{kij} is k_{ij} weight, j=1…m. It reflects to what extent does k_{ij} is related to k_{i0}, $w_{kij} \epsilon$ [0, 1], j=1…m.

After expanding all the paper keywords, arrange the PKS in a descending order according to the number of n-grams of each keyword in it.

3- Annotating paper Sections

This phase determines the paper sections and annotates each of them with its title. A section is a entitled with one of the keywords in the PKS; otherwise it is treated as a zone. "Fuzzy Ontology" and "Term Weighting" are section examples in this paper.

To annotate a certain section, consider its title to apply the following steps to it:

1- Return each pronoun to its referred noun.
2- Remove all stop words.
3- Stem the remaining keywords.

4- If this title includes one of the keywords in PKS, then

i. Put these keywords in a Section Keyword Set, SKS, associating each of them with a weight, w.

$$SKS= \{(s_{lki0}, w_{slki0})\}, i=1\ldots.n, l=1\ldots p$$

Where s_{lki0} is a keyword extracted from the section l title. n is the number of keywords that are extracted from this section's title. p is the number of sections in the paper. w_{slki0} reflects to what extent the keyword k_{i0} represents the section l. The value of $w_{slki0} \epsilon]\ 0,\ 1]$ is calculated through executing the algorithm.

ii. Expand each keyword in the SKS in the given view using the predefined domain fuzzy ontology using the same methodology described previously, so that:

$$SKS= \{(s_lk_{ij}, w_{slkij})\}, i= 1\ldots n, j=0\ldots..m, l=1\ldots p$$

Where, s_lk_{i0} is a keyword extracted from the section s_l's title. w_{slki0} reflects to what extent the keyword k_{i0} represents the section s_l. The value of $w_{slki0} \epsilon]\ 0,\ 1]$ is calculated through this algorithm. p is the number of sections in the document. m is the number of k_{i0}'s expanded keywords. s_lk_{ij} is an expanded keyword for the keyword k_{i0}, j=1…m. w_{slkij} is k_{ij} weight in section s_l. It reflects to what extent k_{ij} is related to k_{i0}, j=1…m.

iii. Arrange the SKS list in a descending order according to the number of n-grams of each element belonging to the SKS.

4- WeightingPhase

This phase calculates the weight of each keyword in the keyword set using Eq. 2 as the summation of its weight in each zone and in each section.

$$w_{k_{in}} = \sum_{l=1}^{y} w_{z_lk_{i0}} + \sum_{l=1}^{p} w_{s_lk_{i0}} \quad (2)$$

Where w_{zlki0} is the weight of the keyword k_i in zone l. w_{slki0} is the weight of the keyword k_i in section l. p is the number of sections in this paper. Y is the number of zones in this paper.

To calculate the weight of a certain keyword in a zone or in a section, the proposed algorithm:

1- Considers only the main sentence of each paragraph in this zone or in this section. A paragraph main sentence is its first two or three lines that represent its main idea.
2- Returns each pronoun in it to its referred noun.
3- Removes all stop words and then stems each of the remaining keywords.
4- Calculates the weight of the keyword k_i that belongs to PKS in a certain zone as in Eq. 3.

$$w_{z_lk_{in}} = w_{z_l} * (freq_{z_lk_{in}} + \sum_{j=1}^{m} w_{k_{ij}} * freq_{z_lk_{ij}}) \quad (3)$$

$$freq_{z_lk_{in}} = \frac{\text{number of occurring } k_{i0} \text{ in zone } z_l}{\text{number of words in the document}} \quad (4)$$

$$freq_{z_lk_{ij}} = \frac{\text{number of occurring } k_{ij} \text{ in zone } z_l}{\text{number of words in the document}} \quad (5)$$

where $w_{z_lk_{in}}$ is the k_{i0}'s weight in the zone z_l. w_{z_l} is the zone z_l's weight. $freq_{z_lk_{in}}$ is the frequency of the occurrence of the keyword k_{i0} in the zone z_l with respect to the number of words in the same zone. $freq_{z_lk_{ij}}$ is the frequency of the occurence of the expanded keyword k_{ij} in the zone z_l with respect to the number of words in the same zone. $w_{k_{ij}}$ is the weight that reflects to what extend the expanded keyword k_{ij} is related to the keyword k_{i0}.

Any keyword in the paper can match with at most one element in the PKS, e.g., consider a PKS= {fuzzy ontology, fuzzy, ontology} and a sentence "fuzzy ontology represents a fuzzy domain" after processing the sentence, it will be "fuzzy ontology represent fuzzy domain". The number of occurence of the word fuzzy ontology in this sentence is 1, zero for the word ontology, and only 1 for the word fuzzy. Also any keyword in the document written in different style (bold, italic, underlined, uppercase…) will take a higher weight.

In the same manner, the weight of the keyword k_i in section l is calculated. A section weight is calculated using Eq. 6 as the ratio between the its number of words to the paper number of words.

$$w_{s_l} = \frac{number\ of\ words\ in\ section\ s_l}{number\ of\ words\ in\ the\ document} \quad (6)$$

4.2 The proposed Algorithm

The algorithm is illustrated as in algorithm1.

Algorithm 1: Term Weighting of research paper in a certain view

Input: research paper in a certain view, a predefined multi-view fuzzy ontology

Output: research paper annotated with a set of weighted keywords in specific view

Steps:

1. Divide the paper into different weighted zones
2. PKS= expand all keywords in the keyword zone according to the given view using the predefined fuzzy ontology
3. Arrange all PKS in descending order according to each keyword n-gram
4. Annotate each zone with the PKS
5. For each zone, calculate the weight of each keyword in PKS
6. For each section
7. annotate it with its title
8. SKS= expand this annotation using the predefined multi-view fuzzy ontology
9. arrange SKS in a descending order with respect to each keyword n-gram
10. calculate the weight of each keyword in SKS
11. End for
12. calculate the weight of each keyword in the keyword zone through summing its value from each zone and each section

5. Tests and Result

The proposed algorithm is tested on 20 research papers. Results are compared with Fernández algorithm. Compared with Fernández algorithm, results show that the proposed algorithm decreases the execution time as the result of working on the paragraph's main sentence instead of working on all the document keywords. Also, the proposed algorithm ranks them more accurately than Fernández algorithm. This accuracy results from using fuzzy ontology to expand the keyword set, annotating each section with its title, dividing the paper into different weighted zones and arranging the expanded list in descending order according to the number of n-grams of each keyword in the keyword set.

Discussion and conclusion

The proposed algorithm's features are compared with others as in table1. These features enhance the resulted weights' accuracy and decrease the execution time. It enhances the resulted weights' accuracy as it uses fuzzy ontology instead of crisp one to expand the keyword set, uses a certain threshold when expanding the document's annotated terms using the predefined fuzzy ontology, arranges the expanded list in descending order according to the number of n-grams of each term in this list, returns each pronoun to its referred noun instead of considering it as stop words, annotates each section with its title, divides the paper into different weighed zones. It reduces the execution time as it only works on the paragraph's main sentences instead of working on all the document words.

Table1: The comparison between the proposed weighting algorithm and others

Weighting algorithm	*Statistical based*	*NLP tools*	*Ontology based*	*Fuzzy Ontology based*	*English Writing rules*
The proposed algorithm	✓	✓	✓	✓	✓
Fernández adapted TF-IDF [8]	✓	✓	✓		
Tar algorithm [9]	✓		✓		

7. References

[1] M. A. A. Leite and I. L. M. Ricarte, "Relating ontologies with a fuzzy information model," Journal Of Knowledge and Information System, pp. 619-651, 2013.

[2] S. Klink, K. Kise, A. Dengel, M. Junker, and S. Agne, "Document Information Retrieval," Digital Document Processing, Springer-Verlag London Limited 2007.

[3] J. Zhai, Y. Liang, Y. Yu and J. Jiang "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce," JOURNAL OF SOFTWARE, VOL. 3, NO. 9, DECEMBER 2008.

[4] Q. T. Tho, S. C. Hui, A. C. M. Fong, T. H. Cao," Automatic Fuzzy Ontology Generation for Semantic Web," IEEE transaction on knowledge and data engineering, Vol. 18, No.6, June 2006.

[5] F. B. Ortenga, M. D. Calvo-Flores, "Managing Vagueness in Ontologies," PHD Dissertation, Granada, October 2008.

[6] Y. Cai, H. F. Leung, "A Formal Model of Fuzzy Ontology with Property Hierarchy and Object Membership," the 27th International Conference on Conceptual Modeling (ER 2008), Vol. 5231, pp.69-82, 2008.

[7] Z. E. Alarab, A. M. Gadallah, H. A. Hefny, "An Enhanced Model For Linguistic-based fuzzy ontology," the 47th Annual Conference on Statistics, computer sciences and operation research, pp. 49-62, 2012.

[8] M. Fernández, I. Cantador , V. López, D. Vallet, P. Castells, E. Motta ," Semantically enhanced Information Retrieval: An ontology-based approach," Web Semantics: Science, Services and Agents on the World Wide Web 9, pp. 432-452, 2011.

[9] H. H. Tar, T. T. S. Nyunt, "Ontology-Based Concept Weighting for Text Documents," International Conference on Information Communication and Management, vol.16, 2011.

[10] P. Soucy, G. W. Mineau, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model," The 19th International Joint Conference on Artificial Intelligence, pp. 1130-1135, 2005.

Computer and Information Technology
10.4028/www.scientific.net/AMM.519-520

Term Weighting: A Multi-View Fuzzy Ontology Based Approach
10.4028/www.scientific.net/AMM.519-520.857