5th International Conference on Control Engineering & Information Technology (CEIT-2022)
Proceedings Book Series –PBS- Vol 5, pp.57-62

Positioning Tags Within Metadata and Available Papers‟ Sections: Is It Valuable for Scientific Papers Categorization?

Djalila Boughareb1, 2, Nadir Farah1, Hamid Seridi2

1 Labged Laboratory-University of Badji Mokhtar Annaba- Algeria
{boughareb, farah}@labged.net
2 LabSTIC Laboratory-University of Guelma- Algeria
{boughareb.djalila, seridi.hamid}@univ-guelma.org

***Abstract*—Nowadays, the reality of getting everything rapidly imposes itself upon getting information. The access to relevant information requires its valuable representation and management. Recent description methods based on tags may have a better impact on the information access task if they are attributed adequately. In order to overcome the categorization problems inherent to the bad document tagging, we propose in this paper a weighting formula based both on the tags' frequencies and positions within the main sections of scientific papers. The results of experiments performed on a CiteULike collection of tagged scientific papers showed that by using the proposed weighting formula we can achieve a significant improvement in text categorization task over classical Term Frequency Inverse Document frequency (TF-IDF) weighting formula with a massive reduction of the computing time.**

Keywords— Automatic text categorization; machine learning; SVM; collaborative tagging;

I. Introduction

Today, with the new generation of information usage techniques on the web, it has been easier to produce and share information and furthermore tagging them without any restrictions. These novelties engendered the continuing increase of different kind of information and especially scientific papers published on the web.

Text categorization or text classification is a branch of text mining field which refers to the process of deriving featured information from text, and classify it to one of predefined classes. In classification process, weighting task consists of attributing values to the terms for measuring their degrees of importance in the document. TF-IDF is considered to be the best term weighting formula [15]. As it favors rare terms and that a collection of tagged documents contains thousands of rare tags with different degree of importance, we thought in a novel tag weighting formula that affects weights to tags according to their significance not only for their number of occurrences.

Based on the following findings: (i) Tags occurring close to the beginning of the document are shown to have a higher key phrases probability [5], [6], [19]. Given that title and abstract summarize the major aspects of the entire paper (purpose, the main idea the paper is talking about, key techniques used, and the major findings including key quantitative results). Therefore, tags occurring in the title and the abstract of a paper are shown to be probably more significant. (ii) Metadata are data that serve to provide context or additional information about the paper, e.g., information about the author, paper‟s context, the year of publication, posted date, posted time, paper priority and the title of the book, journal or conference within which the paper was published that can include tags referring to the topic of the paper. (iii) The contents of scientific papers are not always available online, the publisher‟s website doesn‟t exhibit the complete content of the papers, but only the title and abstract sections‟ are publicly available for researchers. For websites of scientific publications sharing such as CiteULike, Citeseer and Bibsonomy only the title and the abstract and some metadata of the paper plus the associated tags are free access.

In this paper a classification system of tagged scientific papers is proposed, each paper is represented by the title, the abstract and metadata. The system utilizes a new weighting formula to compute term weights based on their frequencies and positions. The main idea is to promote tags and words occurring within the main sections of the paper that are title, abstract and metadata (TAM). The obtained classification results are compared with those obtained using TF-IDF weighting method and the main observation is that a significant improvement gained using TFP (term position and frequency) over TF-IDF.

This paper is organized as follows: Section 2 briefly reviews related work. Section 3 describes the methodology of the proposed approach. Furthermore, section 4, specifies the details of the evaluation. Section 5 discusses obtained results and section 6 concludes the paper and discusses possible directions for future work.

ISSN: 2961-6611

II. Related works

Text classification refers to the process of identifying for each text document the appropriate class. It can be based on supervised learning when the process intervenes humans in defining classes or/and in judging whether it has classified documents correctly. The second sort is known as classification based on unsupervised learning or also clustering, in this case, there is neither predefined class nor an external indication. The third classification type is called classification based on semi-supervised learning, in which some documents are already classified for better learning of the classifier. In classification process, document representation and feature selection are two fundamental tasks [3]; and the most used text representations is the weighted bag of words. The weighting step attributes a value to each term that corresponds to the degree of importance of this term in the document. In this context, TF-IDF is considered to be one of the best term weighting formula [15]. By computing TF-IDF rare terms are favored and considered more important than frequent one, which is not the case for social tags that their low or high frequencies do not reflect their relevancy. In fact, referring to the used collection we found that a lot of rare tags are insignificants. Also, TF-IDF computing requires global statistics about word frequencies in the document collection, so as the collection is bigger as the weight computing step is longer.

Some variations of TF-IDF were proposed in which term frequency (TF) was normalized by relevance frequency [11], [13], by probability based approach [13], by mutual information [13], [8], by odds ratio [13], [8], or by correlation coefficient [13].

The BM25 is one of the most important term weighting formulas, it was developed by Robertson et al., (1995) [4] as a way of probabilistic model that ranks documents considering their lengths and the size of the whole collection besides to the frequency of terms and their distributions within documents.

In our case, since tags do not belong to documents -they are only attributed by users- we used the available sections of the paper to favor significant tags.

More recently, Ren and Sohrab (2013) [18] proposed a class-indexing-based TF-IDF.ICSdF term weighting approach, where ICSdF is the inverse class-space-density frequency. They showed that this new formula gives a positive discrimination to both rare and frequent terms and it outperformed with all the term weighting approaches discussed above.

Some researchers in text classification and indexing proposed to focus on specific sections of the document. Magdy and Darwish (2008) [10] in book searching demonstrated that the use of the title and chapter headings can offer almost the same search effectiveness, such as using the full-text of a book; they also supported the idea that certain metadata elements are more significant in retrieval than others. A similar approach is proposed by Nascimento et al., (2011) [16] where a discriminating weight was assigned to words according to the section where they appear in the papers, they weighted words from title 3 times stronger than words from text body, and words from abstract twice as stronger. Also, Jomsri et al., 2009) [12] used tag, title, and abstract sections to create a paper‟s index, they illustrated that the search engine created based on this index provides better search results as compared to CiteULike search engine, particularly for those documents in the first two ranks which would be considered as the most relevant documents. Moreover, in their work, [9] found that users‟ tags are more appropriate to capture the contents of Web resources than the classical automated contents extraction techniques such as TF-IDF.

The proposed weighting formula differs from previous schemes since it favors neither rare tags such as TFIDF nor popular ones such as TF, but it favors well positioned ones within TAM sections.

III. Methodology

A. Data Set

The data set used in this work is extracted from CiteULike which is a free service for managing and discovering scientific publications that allows the store of resources, their tagging and their sharing in an environment that supports social relations such as friendship and common interest. Each stocked publication is represented through all or some elements of the following list of sections: title, abstract, author (s), editor (s), link to the full text on the editors‟ website, and metadata. CiteULike allows searches by tags.

The extracted collection comprises 2584 tagged publications from 10 computer science fields that are: Collaborative Web (CW), Machine Learning (ML), Information Retrieval (IR), Computer Programming (CP), Computer Architecture (CA), Multi-agents systems (MA), Modeling (M), Semantic Web (SW), Bioinformatics (BI) and Networking and Security (NS). The collection includes 29181 unique words and 24291 unique tags.

This work is interested by title, abstract, meta-properties sections and tags section. Meta-properties section contains generally the following metadata: authors and/or editor‟s names, the title of the book, journal or conference within which the paper was published which can include tags referring to the topic of the paper. For example, the key tag *bioinformatics* appeared in the meta-property of the paper entitled *Stochastic reaction-diffusion simulation with MesoRD*:

```
<meta property="dc:source" content="Bioinformatics, Vol. 21, No.12. (15 June 2005), pp. 2923-2924, doi:10.1093/bioinformatics/bti431" />
```

For the tag field, it comprises tags with their corresponding frequencies. That frequency reflects the popularity of a given tag (how often a tag has been used). The used document collection is described in the table 1.

TABLE I. Description Of The Document Collection

Number of papers	Number of different words	Number of different tags
2584	29181	24291

We denote that the number of tags differs from one paper to another where the most tagged paper has almost 1000 different tags.

B. Tags and Paper's Section Preprocessing

In this step, we extracted from each document, the set of lexical units or terms after eliminating separators: spaces and punctuations‟ signs. In the rest of the paper, we mean by word, each lexical unit extracted from TAM sections. Then, in order to keep only significant tags and words, this works begins by filtering out stop words and non-descriptive tags. Similarly, the treatment of titles, metadata and abstracts (TAM) consists firstly of eliminating stop words, tokenization and measuring the frequency of resulted words. To this end, a filtering list is created manually, including 775 standard English stop-words[1] and non-descriptive tags that are:

- Those combining letters and numbers like (*engs485, file-import-08-05-19, me1021, l1reading*, etc.) and those that contain special characters.
- Those occurred in major papers such as: *paper, approach, method, contribution, summary, model, results, experiments, algorithm, study, framework, figure, table*, etc.
- Subjective Tags which are those expressing user‟s opinions e.g., funny and emotion, e.g., *cool, happy, hate, sad*, etc.

C. Weigh Computing Using TF-IDF

In CiteULike, each tag attributed to a given paper has a value corresponding to its number of occurrence. Indeed, the frequency of tag doesn‟t indicate that it‟s more relevant than tags having lower frequencies which is largely occurring in the used data set. For example, table 2 shows the top frequent tags assigned to the paper „*Why social network are different from other types of network*‟ where we can see that the tag key “*social network*” does not appear in the top-10 popular tags.

TABLE II. EXAMPLE OF TOP-10 POPULAR TAGS OF THE PAPER “*WHY SOCIAL NETWORK ARE DIFFERENT FROM OTHER TYPES OF NETWORK*”

N°	Tags	#Occurrence
1	Link-analysis	30
2	clustering	27
3	webgraph	21
4	ranking	18
5	Link-spam	14
6	Map-reduce	13
7	Link-mining	13
8	graph	11
9	google	11
10	pagerank	10
11	classification	10
12	social-network	9

In this stage, tags and words‟ weights are computed using TF-IDF weighting formula given by the equation (1) [1], [16]. It combines two weighting factors which are: TF (Term Frequency) given by equation (2), that measures the frequency of a term in the document [1] and IDF (Inverse Document frequency) given by equation (3), that measures the frequency of a term in the whole treated corpus [2]. Thus, it‟s required to weigh down the frequent terms TF whereas scale up the rare ones using IDF.

$$TF\text{-}IDF_{i,j} = TF_{i,j} \times IDF_i \qquad (1)$$

Where

$$TF\,(ti, dj) = n\,(ti, dj) / \textstyle\sum_k n\,(tk , dj) \qquad (2)$$

Such as $TF\,(t_i, d_j)$ is the frequency of term t_i in document d_j, $n(t_i, d_j)$ is the number of times t_i occurs in d_j and $\sum_k n\,(t_k, d_j)$ gives the sum of occurrences of each term t_k in d_j.

$$IDF\,(t_i, d_j) = \log\,(|D| / |\{ d_j : t_i \in d_j \}|) \qquad (3)$$

Where $IDF\,(t_i, d_j)$ *is* the inverse frequency of term t_i in document d_j, $|D|$ is the total number of documents in corpus and $|\{d_j : t_i \in d_j\}|$ calculates the number of documents in which the term t_i appears.

D. Weigh Computing Using TFP

By using TF-IDF some infrequent, irrelevant tags will be favored while some other relevant and popular tags will lose their significance. So, we computed a novel weighting formula that takes into account the number of occurrence of the tag in the document and its degree of importance with an assigned favor to tags occurred in TAM sections. The weight of each tag (TFP) is calculated by the equation (4).

$$TFP_{ti} = \log\,(1 + f_{ti}) \times \lambda \qquad (4)$$

Here, fti measures the frequency of tag in the document and λ represents a threshold describing the position of the tag, it takes the following values:

$$\begin{aligned} \lambda &= 0.9 \text{ if the tag occurs in TAM sections} \\ \lambda &= 0.4 \text{ otherwise} \end{aligned} \qquad (5)$$

By according the appropriate weights to tags (weights that estimate their qualities) Carmel et al., [14] obtained significant improvement in search effectiveness. In our case, and as it is illustrated in table 3, for the paper entitled “*Why social network are different from other types of network*” in which the authors compared between social and non-social networks.

The tag *clustering* was the most used tag with 27 times; this tag occurred 2 times in the abstract (A) and zero time in the title (A) and metadata (M) sections. Its weight computed using TF-IDF was 0.0429, and with TFP 0.0726. While, for the tag *link-mining* that did not occur in any sections, it got 0.0251 using TF-IDF and 0.0075 using TFP. So, according to TFP *clustering* is more important than *link-mining*, which has been remarked by consulting the full text of paper when the tag *clustering* occurred 30 times in the whole paper, while the

[1] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop, January 2016

word *link* occurred 4 times only in the whole paper, the word *mining* zero times and the term *link mining* did not occur any time. We can observe clearly that TFP has favored the most relevant tag.

Another example extracted from the data set concerns the paper entitled "*Personalized query expansion for the web*" that proposes to expand short Web queries with terms collected from the user‟s personal information repository. Using TF-IDF we can observe that it has been assigned the same weight to two tags having different relevance that are "*query expansion*" which occurred in the title of the paper and "*word-sense-disambiguation*" that did not occur in any section of the paper. While TFP distinguished between the two tags and gave the high weight to the relevant one.

Similarly, the weights of TAM words are computed using TFP formula where λ is fixed to 0.9.

TABLE III. COMPARISION OF TAG WEIGHTS COMPUTED USING TF-IDF AND TFP

Tags	F	T	A	M	TF-IDF	TFP
Why social network are different from other types of network						
clustering	27	0	2	0	0.0429	0.0726
Link-mining	13	0	0	0	0.0251	0.0075
Personalized query expansion for the web						
word-sense-disambiguation	22	0	0	0	0.0109	0.0038
Query expansion	2	1	0	0	0.0109	0.0153

a. F: frequency, T: title, A: abstract, M: metadata.

E. Classification

Automatic text classification refers to the assignment of a class or a category from a preexisting discovered classes to each text document from the collection. This aims to classify new documents. In this step we firstly grouped similar papers in the same class after expert dealing.

Secondly, we used a machine learning algorithm in order to combine the relevance features and learning a set of rules from a set of papers on the training set. We used SVM classifier that showed its power and effectiveness in the resolution of text classification problems [7], [17], [18], [19]. Then, by using this classification technique, the algorithm first learn the classifier from a training set of documents. The classifier is then applied to the remaining documents in the collection. Finally, we tested the classifier by calculating the probabilities a posteriori of belonging tested documents to different classes.

IV. EVALUATION

In this stage, we verified if the consideration of tags„positions in TAM sections in computing tags‟ weights performs better than computing their weights using TF-IDF. We compared the classification rate by means of expert assessment.

To evaluate the classification rate of the system we computed three evaluation metrics which are precision, recall and F-score.

- *Recall:* It measures the capability of the classification system to detect well classified documents. If all documents are correctly classified recall will take the value 1. We obtain recall for each class using equation (6) by dividing A: the number of document correctly classified into a class, by B: the number of documents belonging really to the same class.

$$Recall = A/B \qquad (6)$$

- *Precision:* It measures the conditional probability that a randomly chosen document were correctly classified. We obtain precision for each class using equation (7) by dividing A: the number of document correctly classified into a class by, C: the total number of documents assigned to the same class.

$$Precision = A/C \qquad (7)$$

- *F-score:* The F-score [21] (harmonic mean of Precision and Recall) is a synthesis indicator which evaluates classification algorithm based on precision and recall. F-score is given using equation (7).

$$F\text{-}Score = ((1+ \beta_2) \times Recall \times Precision) / ((\beta_2 \times Recall) + Precision) \qquad (8)$$

Where $\beta_2 = 1$

V. RESULTS AND DISCUSSION

Using SVM classifier, the classification rate of vector-space based TF-IDF weighting method gave a classification rate of 85% while the classification of vector-space based TFP weighting method gave a classification rate of 98.33 % using the same set of over 850 documents, which means that 835 documents were correctly classified.

Table 4 shows recall, precision and F-score of classification of vector-space based TF-IDF and vector-space based TFP obtained using SVM classifier.

TABLE IV. RECALL, PRECISION AND F-SCORE OF CLASSIFICATION OBTAINED USING SVM CLASSIFIER

Class	TF-IDF			TFP		
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*
CW	0.833	0.833	0.833	1	1	1
ML	1	0.833	0.909	1	0.833	0.909
IR	0.9	1	0.947	1	1	1
CP	1	1	1	1	1	1
CA	1	0.833	0.909	1	1	1
MA	0	0	0	1	1	1
M	1	1	1	0.857	1	0.923
SW	1	0.667	0.8	1	1	1
BI	1	1	0.923	1	1	1
NS	0.455	0.833	0.588	1	1	1
Avg.	0.85	0.85	0.838	0.986	0.983	0.983

We got an average recall of 0.983 with TFP and 0.85 with TF-IDF. For the average precision, we got 0.986 with TFP and 0.85 with TF-IDF, as well as F-score, where we got 0,983 with TFP against 0,838 with TF-IDF. Hence, we got a significant improvement in the three metrics in favor of TFP over TF-IDF.

For the class Modeling (M), TF-IDF gave better results than TFP. Indeed, no or too few tags were assigned to documents of this class.

So, the proposed weighting method (TFP) achieved a significant superiority over classical TF-IDF method and it allowed having considerable improvement in tagged text categorization with a massive reduction of the computing time.

We proved in this work that the computing of word and tag weights based on their position in the title, abstract, and metadata sections helps to effectively represent the content of the paper and to furthermore identify its appropriate class.

VI. Conclusion

In this work, we investigated the effectiveness of the proposed term weighting formula TFP based on words positions in improving text classification task using SVM classifier.

The proposed weighting approach TFP outperformed using SVM classification algorithm over the classical term weighting formula TF-IDF. The experiments were conducted using a *CiteULike* collection including 2584 scientific publications. We also proved that by positioning tags within title, abstract and metadata of papers we can significantly improve the task of scientific papers classification.

In future work, we aim to use TFP formula to filter out irrelevant tags and furthermore improve indexing scientific papers.

References

[1] G. Salton, “Search and retrieval experiments in real-time information retrieval”. IFIP Congress, 2, 1968, pp. 1082-1093.

[2] K. Sparck-Jones, “Experiments in relevance weighting of search terms” Inf. Process. Manage, 15(3), 1979, pp. 133–144.

[3] Y. Yang and J.O. Pederson, “ A comparative study on feature selection in text categorization”, ACM SIGIR Conference, 1995.

[4] S. E. Robertson, S.Walker, M. Hancock-Beaulieu, and M. Gatford. “Okapi at trec-3”, in Text REtrieval Conference (TREC-3), (1995), pages 109–126.

[5] I.H.Witten, GW. Paynter, E. C. Frank Gutwin and CG. Nevill-Manning, “KEA: practical automatic keyphrase extraction”, fourth ACM conference on Digital libraries. Berkeley, California, United States, 1999

[6] PD.Turney, “Learning algorithms for keyphrase extraction”, Information Retrieval, 2, 2000, pp.30-36.

[7] T. Joachims, “A statistical learning model of text classification for support vector machines”, in: Proceedings of the 24th ACM International Conference on Research and Development in Information Retrieval, 2001.

[8] F. Sebastiani, “Machine learning in automated text categorization”, ACM Computing Surveys, 34, 2002 pp.1–47.

[9] X. Li, L. Guo, and Y-E. Zhao, “Tag-based social interest discovery”, Proceedings of the 17th international conference on World Wide Web, 2008, pp.675–684.

[10] W. Magdy, K. Darwish, “Book search:indexing the valuable parts”.In Proceedings of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories, New York, NY, USA, 2008, pp. 53–56.

[11] M. Lan, C.L Su and Y. Lu, “Supervised and traditional term weighting methods for automatic text categorization”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (1), 2009, pp.721–735.

[12] P. Jomsri, S. Sanguansintukul and W. Choochaiwattana, “A Comparison of search engine using “tag title and abstract” with citeulike – An Initial Evaluation,” in the 4th IEEE Int. Conf. for Internet Technology and Secured Transactions (ICITST-2009),United Kingdom,2009.

[13] Y. Liu, H. Loh and A. Sun, “Imbalanced text classification: a term weighting approach”, Expert Systems with Applications 36, 2009, pp. 690–701.

[14] D. Carmel, H. Roitman and E. Yom-Tov, “Social bookmark weighting for search and recommendation”, The VLDB Journal, 19 (6), 2010, pp. 761–775.

[15] S. Flora and T. Agus, “Experiments in term weighting for novelty mining”, Expert Systems with Applications, 38, 2011, pp. 14094–14101.

[16] C. Nascimento, A. H. Laender, A. S. da Silva, and M. A. Gonçalves, “A source independent framework for research paper recommendation,” in Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, 2011, pp. 297–306.

[17] C.H.Wan, L.H. Lee, R. Rajkumar and D. Isa, “A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine”, Expert Systems with Applications 39, 2012, pp.11880–11888

[18] F. Ren and M.G. Sohrab, “Class-indexing-based term weighting for automatic text classification”, Information Sciences, 236, 2013, pp.109–125.

[19] A. Joorabshi, M. English and A.E. Mahdi, “Automatic mapping of user tags to wikipedia concepts: the case of a q&a website– stackoverflow”, Journal of Information Science, 2014, pp. 1–15.

[20] M-A. Siddiqui “An empirical evaluation of text classification and feature selection methods”, Artificial Intelligence Research , 5(2), 2016.

[21] K. Van Rijsbergen, “Information Retrieval”, (2nd Ed.) Butterworths, London. www.dcs.gla.ac.uk/Keith/Preface.html