VSMs with K-Nearest Neighbour to Categorise Arabic Text Data

Fadi Thabtah, Wa'el Musa Hadi, Gaith Al-shammare

***Abstract*—Text categorisation is a popular problem that has been studied extensively in the last four decades. This paper investigates different variations of vector space models (VSMs) and term weighting approaches using KNN algorithm. The base of our comparison in the experiments we conduct is the F1 evaluation measure. The Experimental results against different Arabic text categorisation data sets provide evidence that Dice and Jaccard Coefficient outperform the Cosine Coefficient approach with regards to F1 results, and the Dice-based TF.IDF achieves the highest average scores.**

***Index Terms*—Data sets, Data mining, Text categorisation, Term weighting, VSM.**

I. Introduction

Text categorisation (TC) is one of the important tasks in information retrieval (IR) and data mining [14]. This is because of the significance of natural language text, the huge amount of text stored on the internet, and the available information libraries and document corpus. Further, TC importance rises up since it concerns with natural language text processing and classification using different techniques, in which it makes the retrieval and other text manipulation processes easy to execute.

Many TC approaches from data mining and machine learning exist such as: decision trees [9], Support Vector Machine (SVM) [4], rule induction [8], and Neural Network [19]. The goal of this paper is to present and compare results obtained against Arabic text collections using K-Nearest Neighbour algorithm. Particularly, three different experimental runs of the KNN algorithm (Cosine, Dice, Jaccard) on the Arabic data sets we consider are performed.

Generally, TC based on text similarity goes through two steps: Similarity measurement and classification assignment. Term weighting is one of the known concepts in TC, which can be defined as a factor given to a term in order to reflect the importance of that term. There are many term weighting approaches, including, inverse document frequency (IDF), weighted inverse document frequency (WIDF) and inverse term frequency (ITF) [16]. IDF and WIDF focus on terms occurrences inside a text corpus. WIDF distinguishes between two terms that have different occurrences, whereas, IDF treats both terms equally. In this paper, we compare different variations of VSMs with KNN [20] algorithm using IDF and WIDF. The base of our comparison between the different implementations of the KNN is the F1 measure [18]. In other words, we want to determine the best VSM, which if merged with KNN produces good F1 results. To the best of the author's knowledge, there are no comparisons which have been conducted against Arabic language data collections using VSM.

The organisation of this paper is as follows, related works are discussed in Section 2. TC problem is described in Section 3. In Section 4, experiment results are explained, and finally conclusions and future works are given in Section 5.

Fadi Thabtah is in the MIS Dept, Philadelphia University, Amman, Jordan Email: ffayez@philadelphia.edu.jo

Wa'el Musa Hadi is in the CIS Dept, Arab Academy for Banking and Financial Sciences, Jordan. Email: whadi81@students.aabfs.org

Gaith Al-shammare is in the Software Engineering Dept, Philadelphia University, Amman, Jordan. Email: gaithshammare@yahoo.com

II. Related Works

Since TC stands at the cross junction to modern IR and machine learning, several research papers have focused on it but each of which has concentrated on one or more issues related to such task. There are few previous works on Arabic TC. For instance, [7] compared between Manhattan distance and Dice measures using N-gram frequency statistical technique against Arabic data sets collected from several online Arabic newspaper websites. The results showed that N-gram using Dice measure outperformed Manhattan distance. The author's of [13] presented results using statistical methods such as maximum entropy to cluster Arabic news articles; the results derived by these methods were promising without morphological analysis. In [6], Naïve Bayes was applied to classify Arabic web data, the results showed that the average accuracy was 68.78%. [2] used Maximum Entropy for TC on Arabic data sets, the results revealed that the average F-measure increased from 68.13% to 80.41% using preprocessing techniques (normalisation, stopwords removal, and stemming). Finally, the algorithm developed by [1] has outperformed other presented text classification algorithms, i.e. [6], [2], [13], and Sakhr's Categorizer [10] with regards to F-measure results.

III. Text Categorisation Problem

TC is one of the well studied problems in data mining and IR. Given a large quantity of documents in a data set where each document is associated with its corresponding categories. The categorisation involves building a model from classified documents, in order to classify previously unseen documents as accurately as possible. TC problem can be defined according to [14] as follows: let G denotes the collection of categories which contain $\{g_1,g_2,\ldots..,g_n\}$, let D denotes the collection of documents $\{d_1,d_2,\ldots,d_j\}$. Also, let R denotes the set of classifiers for $D \times G \to \{T, F\}$, each pair (d_j,g_n) where document $d_j \varepsilon D$ is assigned a Boolean value i.e. T or F, where T indicates a document that belongs to g_n, while a value of F indicates a document not belonging to category g_n.

A. Term Weighting Measures

Term weighting is one of the important issues in TC, which has been widely investigated in IR [12]. Term weighting corresponds to a value given to a term to reflect the importance of that term in a document in order to improve the classification performance. There are many term weighting approaches, including, IDF, TF.IDF, WIDF, ITF, log(1+TF), and so on. Table I explains some of these approaches.

Where

N: The total number of the given documents.

n: The number of documents, which contain a specified term.

B. Similarity Measurements

There are several well-known similarity techniques, such as: VSM [11], and Probabilistic Model [16]. The VSM is one of the popular models in IR systems. As indicated by its name, it is based on the concept of t-dimensional vectors; more details are given in [11]. In this paper we focus on VSM by adapting different measures: Cosine (shown in equation (1)), Jaccard (shown in equation (2)), and Dice (shown in equation (3)).

$$Sim(Vi,Vj) = \frac{\sum_{k=1}^{m}(W_{ik} \times W_{jk})}{\sqrt{\sum_{k=1}^{m} W^2{}_{ik} \times \sum_{k=1}^{m} W^2{}_{jk}}} \quad (1)$$

$$Sim(Vi,Vj) = \frac{\sum_{k=1}^{m}(W_{ik} \times W_{jk})}{\sum_{k=1}^{m} W^2{}_{ik} + \sum_{k=1}^{m} W^2{}_{jk} - \sum_{k=1}^{m}(W_{ik} \times W_{jk})} \quad (2)$$

$$Sim(Vi,Vj) = \frac{2\sum_{k=1}^{m}(W_{ik} \times W_{jk})}{\sum_{k=1}^{m} W^2{}_{ik} + \sum_{k=1}^{m} W^2{}_{jk}} \quad (3)$$

Where V_i is the document, V_j is the incoming document, W_{ik} corresponds to the weight of the k-th element of the term vector V_i, i.e. pre-categorised documents, and W_{jk} is the weight of K-th element of the term vector V_j i.e. incoming text. The greater the value of $Sim(V_i,V_j)$, the more similar these two texts are.

C. Classification Assignment

There are many approaches to assign categories to incoming text such as (SVM) [4], Neural Network [19] and k-nearest neighbor (KNN) [20]. In our paper, we implemented text-to-text comparison (TTC), which is also known as the KNN [20]. KNN is a statistical classification approach, which has been intensively studied in pattern recognition over four decades. KNN has been successfully applied to TC problem, i.e. [21], [20], and showed promising results if compared with other statistical approaches such as Baysian based Network [17].

TABLE I
TERM WEIGHTING APPROACHES

Term weighting	Equation	Description
TF	TF(d,t)	Term Frequency
IDF	log (N/n)	Inverse Term Frequency
TF.IDF	TF (t).IDF (t)	Combine Term Frequency with Inverse Term Frequency
WIDF	$\frac{TF(d,t)}{\sum_{i \in D} TF(i,t)}$	Weighted Inverse Term Frequency
ITF	1-1/(1+TF)	Inverse Term Frequency
LOGTF	Log(1+TF)	Logarithmic Frequency

IV. Experiment Results

The data used in our experiments are data sets collected from online Arabic newspapers including *Al-Jazeera*, *Al-Nahar*, *Al-hayat*, *Al-Ahram*, and *Al-Dostor*. Arabic text is different than English one since Arabic language is highly inflectional and derivational language which makes monophonical analysis a complex task. Also, in Arabic script, some of the vowels are represented by diacritics which usually left out in the text and it does use capitalisation for proper nouns that creates ambiguity in the text [3].

Three TC techniques based on vector model similarity (Cosine, Jaccard, and Dice) have been compared in term of F1 measure, which is shown in equation (4). These methods use the same strategy to classify incoming text i.e. KNN. We have several options to construct a text classification method; we compared techniques using different term weighting IDF,

TABLE II
F1 RESULTS FOR THE ARABIC TEXT CATEGORISATION DATA

Category	Cosine				Dice				Jaccard			
	TF.IDF	WIDF	ITF	Log(1+TF)	TF.IDF	WIDF	ITF	Log(1+TF)	TF.IDF	WIDF	ITF	Log(1+TF)
Ariculture	95.24	82.86	93.75	96.77	96.77	88.24	93.75	95.24	96.77	88.24	93.75	95.24
Art	80.77	59.09	86.27	84.62	87.27	63.83	88.46	82.14	87.27	63.83	88.46	82.14
Economy	90.00	75.00	89.29	93.10	96.55	76.92	90.00	86.21	96.55	76.92	90.00	86.21
Health	93.94	81.25	97.06	95.52	97.06	89.55	92.31	83.87	97.06	89.55	92.31	83.87
Politics	94.74	79.31	94.92	96.67	96.55	84.75	90.91	86.21	96.55	84.75	90.91	86.21
Science	90.91	78.13	84.85	89.23	95.24	82.76	82.35	77.61	95.24	82.76	82.35	77.61
Average	**90.93**	**75.94**	**91.02**	**92.65**	**94.91**	**81.01**	**89.63**	**85.21**	**94.91**	**81.01**	**89.63**	**85.21**

WIDF, ITF and log (1+tf).KNN. All of the experiments were implemented using VB.NET on 2.8 Pentium IV machine with 256 RAM.

$$F1 = \frac{2 * \Pr ecision * \operatorname{Re} call}{\operatorname{Re} call + \Pr ecision} \quad (4)$$

Precision and Recall shown in equations (5,6) are known measures used to evaluate IR techniques.

$$\Pr ecision = \frac{TP}{(TP + FP)} * 100 \quad (5)$$

$$\operatorname{Re} call = \frac{TP}{(TP + FN)} * 100 \quad (6)$$

For Group of documents g_i, the FP is the number of documents not belonging to group g_i that have been incorrectly categorised as group g_i ; the TP is the number of documents belonging to group g_i that have been correctly categorised as group g_i ; FN is the number of documents belonging to group g_i that have been incorrectly categorised as group g_k.

Table II gives the F1 results generated by the three categorisers (Cosine, Dice and Jaccard) against six Arabic data sets; where in each data set we consider 70% of documents arbitrary for training, and 30% for testing. *K* parameter in the KNN algorithm was set to 11.

After analysing Table II, we discovered that there is consistency between Dice based on TF.IDF and Jaccard based on TF.IDF algorithm in which both of them outperformed Cosine based TF.IDF, Cosine based WIDF, Cosine based ITF, Cosine based log(1+tf), Dice based WIDF, Dice based ITF, Dice based log(1+tf), Jaccard based WIDF, Jaccard based ITF, and Jaccard based log(1+tf). Particularly, Dice based TF.IDF outperformed Dice based WIDF, Dice based ITF, Dice based log(1+tf), Jaccard based WIDF, Jaccard based ITF, Jaccard based log(1+tf), Cosine based TF.IDF, Cosine based WIDF, Cosine based ITF, and Cosine based log(1+tf) on 6,5,6,6,5,6,6,6,5, and 4 data sets, respectively.

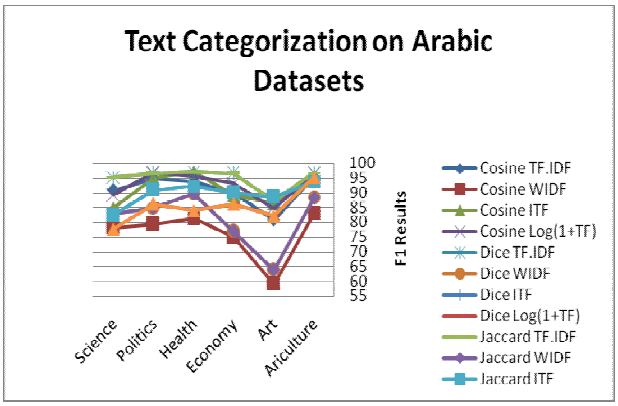

Fig 1. F1 measure results

According to Fig 1, there are similarities between Dice and Jaccard methods, in which both produced similar F1 results for the same term weighting. Cosine based WIDF has the least F1 results on the data sets we consider. Dice TF.IDF and Jaccard TF.IDF achieved the highest scores on Health data sets; Cosine WIDF achieved the least score against the Art data sets. WIDF term weighting approach has the lowest F1 results with the same similarity function.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we investigated different variations of VSM using KNN algorithm, these variations are Cosine coefficient, Dice coefficient and Jacaard coefficient, using different term weighting approaches. The average F1 results obtained against six Arabic data sets indicated that Dice based TF.IDF and Jaccard based TF.IDF outperformed Cosine based TF.IDF, Cosine based WIDF, Cosine based ITF, Cosine based log(1+tf), Dice based WIDF, Dice based ITF, Dice based log(1+tf), Jaccard based WIDF, Jaccard based ITF, and Jaccard based log(1+tf). In near future, we intend to propose a new multi-label classification approach based on association rule for the TC problem, and we intend to build larger Arabic Language TC data sets.

REFERENCES

[1] A. El-Halees, "Mining Arabic Association Rules for Text Classification" In the proceedings of the first international conference on Mathematical Sciences. Al-Azhar University of Gaza, Palestine, 15 -17 (2006). To be appear.

[2] A. El-Halees, "Arabic Text Classification Using Maximum Entropy" The Islamic University Journal (Series of Natural Studies and Engineering)Vol. 15, No.1, pp 157-167, 2007
[3] B. Hammo, H. Abu-Salem, S. Lytinen, and M. Evens, "QARAB: A Question Answering System to Support the Arabic Language". Workshop on Computational Approaches to Semitic Languages. ACL 2002, Philadelphia, PA, July. pp. 55-65.
[4] T. Joachims, "Text Categorisation with Support Vector Machines: Learning with Many Relevant Features" Proceedings of the European Conference on Machine Learning (ECML), (pp. 173-142). Berlin, 1998, Springer.
[5] M. Junker, R. Hoch, and A .Dengel, "On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy," in Proceedings of the Fifth International Conference on Document Analysis and Recognition. 1999.
[6] M. El-Kourdi, A. Bensaid, T. Rachidi, "Automatic Arabic Document Categorisation Based on the Naïve Bayes Algorithm," 20th International Conference on Computational Linguistics . August 28th. Geneva (2004).
[7] L. Khreisat, "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study," DMIN 2006: 78-82
[8] I. Moulinier, G. Raskinis, J. Ganascia, "Text categorization: a symbolic approach," Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval,1996.
[9] J. Quinlan, "C4.5: Programs for machine learning," San Mateo, CA: Morgan Kaufmann.
[10] Sakhr software company's website: www.sakhrsoft.com, 2004.
[11] G. Salton, A. Wong and C.S. Yang. "A VSM for Automatic Indexing," Communications of the ACM, 18, 613-620. 1975.
[12] G. Salton, and M.J McGill, "Introduction to Modern Information Retrieval" .McGill-Hill,1983.
[13] H. Sawaf, J. Zaplo and H. Ney, "Statistical Classification Methods for Arabic News Articles". Arabic Natural Language Processing, Workshop on the ACL'2001. Toulouse, France, July.
[14] F. Sebastiani, "Text categorization," In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, 2005, pp. 109—129
[15] F. Thabtah, P. Cowling, and Y. Peng, "MMAC: A new multi-class, multi-label associative classification approach," Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM '04), (pp. 217-224). Brighton, UK.
[16] T. Tokunaga, M. Iwayama, "Text Categorisation Based on Weighted Inverse Document Frequency," 1994, Department of Computer Science, Tokyo Institute of Technology: Tokyo, Japan.
[17] K. Tzeras, S. Hartman, "Automatic indexing based on bayesian inference networks," Proceedings of the 16th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93), (pp. 22-34, 1993).
[18] C. Van Rijsbergan, "Information retrieval," Buttersmiths, London, 2nd Edition, 1979.
[19] E. Wiener, J. O. Pedersen, and A. S. Weigend, "A neural network approach to topic spotting," Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), pp. 317-332, Las Vegas, Nevada, 1995.
[20] Y. Yang, "An evaluation of statistical approaches to text categorization," Journal of Information Retrieval, 1(1/2):67-88, 1999.
[21] Y. Yang, X. Liu,"A re-examination of text categorisation methods," Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), (pp. 42-49), 1999.