ISSN 2250-0987
Boumedyen et al, UNIASCIT, Vol 1 (3), 2011, 90-94

CLUSTERING THE ARABIC DOCUMENTS (CAD)

Boumedyen #1, Victor *2

Information System, University of Nizwa, Information System, SPIRAN University Barkat-ul- Mauz, Nizwa, Sultanate of Oman, St.Petersburg, Russia

* *Information System, SPIRAN University*

***Abstract*— This work present new clustering techniques to discover when two documents are similar. The proposed techniques yields a grouping of the documents into clusters of closely related items, this clustering can help solve the problems of document duplication. The new clustering techniques are based on LIPNS ,SAMA ,Text Normalization, DNSA and , NADST techniques respectively. We believe that This work will provide new functionality for dealing with huge amount of information on the web .It allows search engines to better present results to their clients and to measure the similarity of any two objects or cluster the sets of similar documents from a large corpus. The new clustering technique was implemented on Arabic documents. The main source of dataset was collected from Latifa Al-Sulaiti's homepage provides its fourteen classifications: Autobiography , Short Stories , Children's Stories, Economics, Education, Health and Medicine , Interviews , Politics , Recipes , Religion Science, Sociology , Sports , Tourist and Travel .The Experimental results shows that our CAD system outperforms other common systems**

***Keywords*— Clustering, Text Normalization, Text Similarity, Stemming**

I. INTRODUCTION

[1] [2] [3] European languages such as French, German, and Spanish and in Asian languages such as Chinese and Japanese. Nevertheless, in Arabic language there is little ongoing research in automatic Arabic document classification [4].

II. RELATED WORK IN ARABIC TEXT CLASSIFICATION

Text organization/categorization is a problem in information science. The mission is to allocate an electronic document to one or more categories, based on its contents [5].

[6][7][8] used statistical classification methods such as maximum entropy to classify and cluster news articles. In addition, [9] El-Halees (2006) described a method based on association rules to classify Arabic documents Other works can be fond in [10] [11] [12][13] [14][15][16][17].

.

III. SYSTEM DESCRIPTION

The documents divided in two datasets; one for training and one for testing. Let training data set = {d1,d2,…,dg}, where g documents are used as examples for the classifier, and must contain sufficient number of positive examples for all the categories involved [18] . The testing data set {dg+1,dg+2,…,dn} used to test the classifier effectiveness . Generally, Classification goes through three main steps: Data pre-processing, text classification and evaluation. Data pre-processing phase is to make the text documents suitable to train the classifier. Then, the text classifier is constructed and tuned using a text learning approach against from the training data set. Finally, the text classifier gets evaluated by some evaluation measures i.e. recall, precision, etc. Our proposed algorithm CAD attempts to attain a better understanding and elaboration of Arabic text classification techniques. CAD presents a different way from other available methods. The internet was used to compile the dataset. Latifa Al-Sulaiti's homepage provides its fourteen classifications: Autobiography , Short Stories , Children's Stories, Economics, Education, Health and Medicine , Interviews , Politics , Recipes , Religion Science, Sociology , Sports , Tourist and Travel . Text Similarity LIPNS [19] , Stemming Techniques SAMA1 [20] , texts normalization [21] , DNSA [22] , and NADST techniques [23] , respectively were used in this work .

For evaluating our algorithm, Latifa Al-Sulaiti's dataset was compiled into one corpus, and performing the following operation on the compiled corpus:

- building document/term matrix from Arabic corpus
- calculate length (words count)of documents
- divides document length into five parts
- extract most frequently terms (five) for each part
- repeat steps from 2 to 4 for all document
- calculate similarity between two documents using LIPNS techniques [19]
- groups similar documents into category

Example:

Table 1
Context of two text documents D1 and D2

Text from D1	Text from D2
يعتبر التصحر مشكلة عالمية تعانى منها العديد من البلدان في كافة أنحاء العالم. ويعرف علي أنه تناقص في قدرة الإنتاج البيولوجى للأرض أو تدهور خصوبة الأراضي المنتجة بالمعدل الذي يكسبها ظروف تشبه الأحوال المناخية الصحراوية. لذلك فإن التصحر يؤدى إلي انخفاض إنتاج الحياة النباتية، ولقد زاد مجموع المساحات المتصحرة في العالم وخصوصا في الوطن العربي ونجد ان ارتفاع درجة الحرارة وقلة الأمطار أو ندرتها تساعد علي التصحر	ونجد أن العوامل التى تساهم في ظاهرة التصحر هى التغيرات المناخية:ارتفاع درجة الحرارة وقلة الأمطار أو ندرتها تساعد علي سرعة التبخر وتراكم الأملاح في الأراضي المزروعة (فترات الجفاف). - كما أن السيول تجرف التربة وتقتلع المحاصيل مما يهدد خصوبة التربة. - زحف الكثبان الرملية التى تغطى الحرث والزرع بفعل الرياح. - ارتفاع منسوب المياه الجوفية. - الزراعة التى تعتمد علي الأمطار.ان التصحر مشكلة عالمية في العديد من البلدان العربية و في كافة أنحاء العالم. ويعرف علي أنه تناقص في قدرة الإنتاج البيولوجى للأرض أو تدهور خصوبة الأراضي المنتجة بالمعدل الذي يكسبها ظروف تشبه الأحوال المناخية الصحراوية. لذلك فإن التصحر يؤدى إلي انخفاض إنتاج الحياة النباتية، ولقد بلغ مجموع المساحات المتصحرة في العالم حوالى 46 مليون كيلومتر مربع يخص الوطن العربى منها حوالى 13 مليون كيلومتر مربع أى حوالى 28 % من جملة المناطق المتصحرة في العالم

After performing SAMA1 stemming on D1 and D2 the obtained results are:

Table 2
Context of two text documents D1 and D2 after using SAMA1 stemming .

Text of D1 after Stemming	Text of D2 after stemming
صحر شكل علم عنى عدد بلد كفى أنحاء علم عرف نقص قدر نتج بيولوجى ارض دهور خصب ارض نتج عدل كسب ظر ف شبه حول منخ صحر صحر خفض نتج حياة نبت زاد جمع مسح صحر علم خصص وطن عرب رفع درج حرارة قلة مطر ندرت سعد صحر	عمل سهم ظهر صحر غير منخ رفع درج حرارة قلة مطر ندر سعد سرع بخر ركم ملح ارض زرع سيل جرف ترب قلع حصل هدد خصب ترب زحف كثب رمل غطى حرث زرع فعل رياح رفع نسب مياه جوف زرع عمد مطر صحر شكل علم عدد بلد عرب كفى أنحاء علم نقص قدر نتج بيولوجى ارض دهور خصب ارض نتج عدل كسب ظر ف شبه حول منخ صحر صحر خفض نتج حياة نبت بلغ مجموع مسح صحر علم حول مليون كيلومتر ربع خصص وطن عرب حول مليون كيلومتر ربع حول جمل نطق صحر علم

A. Documents Statistics'

Len (D1) = 47 terms
Len(D2) = 93 terms

Dividing Len(D) by 5 that is:
D1= 47 \ 5 = 9
D2 = 93 \ 5 = 18

Table 3
Context of two text documents D1 and D2 after Groups process

Group(w9)	Groups of D1	Group(w18)	Groups of D2
1	صحر شكل علم عنى عدد	1	عمل سهم ظهر صحر عدد
2	عرف نقص قدر نتج ارض	2	زرع سيل جرف ترب قلع
3	نتج عدل كسب ظر ف صحر	3	نسب مياه جوف زرع عمد
4	خفض نتج حياة نبت جوف	4	ارض خصب نتج صحر خفض
5	خصص وطن عرب رفع درجة	5	جمع مسح صحر علم مليون

After Using LIPN algorithm the results are:

Table 4
Matrix of D1 and D2 using LIPNS

D!\D2	عمل	سهم	ظهر	صحر	عدد	زرع	سيل	جرف	ترب	قلع	نسب	مياه	جوف	زرع	عمد	ارض	خصب	نتج	صحر	خفض	جمع	مسح	صحر	علم
صحر	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
شكل	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
علم	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
عنى	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
عدد	0	0	0		1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
عرف	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
نقص	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
قدر	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
نتج	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ارض	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
نتج	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
عدل	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
كسب	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ظرف	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
صحر	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
خفض	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
نتج	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
حياة	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
زاد	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
خصص	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
وطن	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
عرب	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
رفع	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
درجة	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
جوف	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Sum(1) = 15

Min(len (Groupw(dj)) , len(Groupw(dj+1))) = 9

Similarity(Dj,Dj+1)= Sum(1) \ Min(len (Groupw(dj)) , len(Groupw(dj+1))

Similarity(Dj,Dj+1) = 15 \ 9

Similarity(Dj,Dj+1) = 1.6666

If Similarity(Dj,Dj+1) > = 1 then

Dj and Dj+1 are similar
Store them in one category

Else
Compare Dj with Dj+ 2 etc….

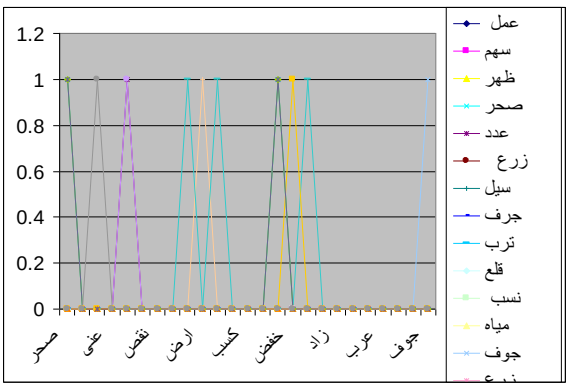

Figure 1 : Result of LIPNS for D1,D2 matrix in table 4

iv. Experimental Results

We select five classifiers systems and compare their performance in the point of precision, recall. These accuracy measures are defined as follows:

Precision = number of correct classes found / number of classes found

Recall = number of classes found / number of correct classes

Table 5
Performance of CAD systems compared with other systems.

Classifier System	Recall	Precision
K-Nearest Neighbor	66.54	66.52
Naive Bayesian	74.49	74.43
Concept Vector-based	80.77	80.74
Sakhr's Categorizer	72.58	46.66
El-Kourdi	70.66	65.77
CAD	89.34	88.34

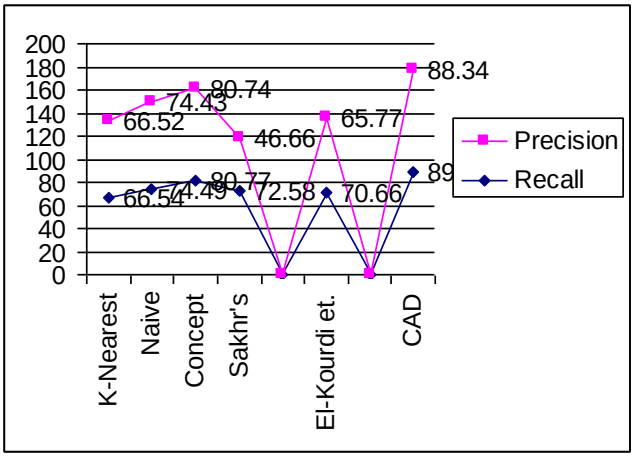

Figure 2 Performance of CAD systems compared with other systems.

From table 5 we can see that our developed system CAD has the best precision and recall The CAD outperform other system. We recorded the performance for each class of the 14 categories Table 6 and figure 2 illustrates the results.

Table 6
The Performance of CAD in each of the domain category.

Category	Recall	Precision
Autobiography	77.38	77.30
Short Stories	70.33	69.98
Children's Stories	88.96	78.67
Economics	78.12	87.42
Education	87.65	82.32
Health and Medicine	70.44	68.21
Interviews	92.45	90.22
Politics	71.83	71.11
Recipes	94.65	92.91
Religion	99.21	98.54
Science	80.52	90.42
Sociology	97.17	90.14
Sports	99.99	99.22
Tourist and Travel	95.86	95.31

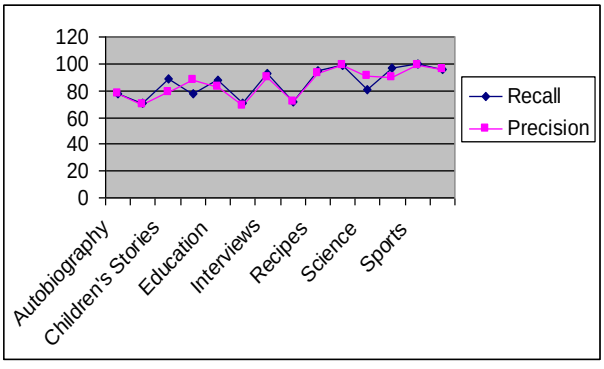

Figure4.2: The Performance of CAD in each of the domain category

v. Conclusion

In this paper We proposed a new classification approach CAD based on LIPNS. We tested the CAD system using real data collected from Latifa Al-Sulaiti's homepage provides its fourteen classifications: Autobiography , Short Stories , Children's Stories, Economics, Education, Health and Medicine , Interviews , Politics , Recipes , Religion Science, Sociology , Sports , Tourist and Travel In our experiments, we computed recall (the percentage of the total documents for the given topic that are correctly classified) and precision (the percentage of predicted document for the given topic that are correctly (classified) which are generally accepted ways of measuring system's performance in this field . We tested our system and compared its overall performance with others exiting systems. The results are recorded in Table 5 . From this table we can notice that CAD system outperform other systems.

REFERENCES

[1] F.Sebastiani, "A Tutorial on Automated Text Categorization," In Proceedings of the ASAI-99, 1st Argentinian Symposium on Artificial Intelligence,
1999. pp. 7-35.

[2] Inderjit S. Dhillon and Dharmendra S. Modha, "Concept Decompositions for Large Sparse Text Data using Clustering,", Machine Learning, vol. 42:1, pp. 143-175, January, 2001.

[3] Robb, D., Text Mining Tools Take on Unstructured Information. Computerworld, 21 June (2004).

[4] Ciravegna, F., Gilardoni, L., Lavelli, A., Ferraro, M., Mana , N., Mazza, S., Matiasek, J., Black, W., Rinaldi, F., Flexible Text Classification for Financial Applications: the FACILE System. In Proceedings of PAIS-2000, Prestigious Applications of Intelligent Systems sub-conference of ECAI2000. (2000)
[5] http://en.wikipedia.org/wiki/Document_classification

[6] Tobias Sche er , Stefan Wrobel "Text Classi cation Beyond the Bag-of-Words Representation ", University of Magdeburg, FIN/IWS, Universit atsplatz 2, 39106 Magdeburg, Germany .

[7] El-Kourdi, M., Bensaid, A., Rachidi, T., Automatic Arabic Document
Categorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics . August 28th. Geneva (2004).

[8] Sawaf, H., Zaplo, J., Ney, H., Statistical Classification Methods for Arabic News Articles. Arabic Natural Language Processing, Workshop on the ACL'2001. Toulouse, France, July (2001).

[9] El-Halees A., 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).

[3] Sauban, M. , Pfahringer, B, Text Categorization Using Document Profiling. Principles of Data Mining and Knowledge Discovery. (2003)

[4] Yang, Y., Slattery, S., Ghani, R., A Study of approaches to hypertext Categorization. Journal of Intelligent Information Systems Vol. 18 p. 219-214 (2002).

[6] Sebastiani, F., Machine learning in automated text categorization, ACM Computing Surveys (CSUR) Vol. 34 , Issue 1. P:1 - 47 (2002) [5] Lewis, D., Naïve (Bayes) at forty: The Independent Assumption in Information Retrieval. In Machine Learning: ECML-98, 10th European Conference on Machine Learning. p 4-15 (1998).

[8] Joachims, T., Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of ECML-98, 10th European Conference on Machine Learning. Pages 137-142. (1998).

[9] Yang, Y., An Evaluation of Statistical Approaches to Text Categorization . Journal of Information Retrieval . (1999).

[10] Nigam, K., Lafferty, J., McCallum, A., Using Maximum Entropy for Text Classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67. (1999).

[11] Hammo, B., Abu-Salem, H., Lytinen, S., Evens, M., QARAB: A Question Answering System to Support the Arabic Language. Workshop on Computational Approaches to Semitic Languages. ACL 2002, Philadelphia, PA, July. p 55-65 (2002).

[12] Y. Yang, T. Ault, T. Pierce, and C. W. Lattimer, "Improving text categorization methods for event tracking," Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00), pp65-72, 2000.

[13] Y. Yang and X. Liu, "A re-examination of text categorization methods," Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pp 42--49), 1999.

[14] Kjersti ,A. Eikvil, L., Text categorization - A survey. Report No. 941,June, 19 (1999).166 Alaa M. El-Halees.

[15] Peng, F., Huang, X., Schuurmans, D., Wang, S., Text Classification in Asian Languages without Word Segmentation. In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003), Association for Computational Linguistics, July 7, Sapporo, Japan. (2003).
[16] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp412-420, 1997.

[17] M. Craven, D. Dipasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to Construct Knowledge Bases from the World Wide Web," Preprint submitted to Artificial Intelligence, Jan, 2000.

[18] Fadi Thabtah, Mohammad Ali H. Eljinini, Mannam Zamzeer ," Naïve Bayesian Based on Chi Square to Categorize Arabic Data " ,Jordan Philadelphia University,2009 .

[19] Boumedyen A.N. Shannaq , "Language Independent Product Name Search (LIPNS) " , First IEEE International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2008), VSB- Technical University of Ostrava, Czech Republic August 4- 6, 2008.

[20] Boumedyen Shannaq, S.V. Kuleshov," Super Arabic morphological analyzer (SAMA1) " St. Petersburg institute for Informatics and Automation of Russian RAS ,Academy of Sciences, 199178, Russia ,VAX UDC 003.9, information-measuring and operating systems Journal http://www.radiotec.ru/catalog.php?cat=jr **/ /**.(2009)

[21] Boumedyen A.N. Shannaq , " Diagonal Name Search For Arabic (DNSA) " , First E-Technologies and Environment Conference (ETEC08) 15-16 April,2008 Sohar , Oman.

[22] P.P.Kokorin , Boumedyen A.N. Shannaq ,"Methods of texts normalization and ontological clustering of texts" , St. Petersburg institute for Informatics and Automation of Russian RAS ,Academy of Sciences, 199178, Russia ,VAX UDC ,information-measuring and operating systems Journal, http://www.radiotec.ru/catalog.php?cat=jr.(2010)

[23] Boumedyen ,The New Arabic Document Summarization techniques (NADST) , MECIT ,2011.