An Automatic Construction of Arabic Similarity Thesaurus

Abdulaziz Al-Qabbany*, AbdulMalik Al-Salman** and Abdulrahman Almuhareb*
* King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia. Email: {qabbany, muhareb}@kacst.edu.sa
** Computer Science Dept., King Saud University, Riyadh, Saudi Arabia. Email: salman@ksu.edu.sa

Abstract*— In this paper, we propose an improvement to the similarity thesaurus construction method used for query expansion in information retrieval. Our proposed method shows an improvement of about 3.3% over its predecessor method. The method was used for constructing an Arabic thesaurus and can be used with other languages as well.

***Keywords*— Automatic Thesaurus Construction; Automatic Thesaurus Generation; Query Expansion**

I. INTRODUCTION

The main objective of information retrieval is to effectively retrieve user's required information items from large collections. Effective information retrieval systems provide the user with all relevant items to his/ her query and only those items. These items can be of different formats such as documents, pictures and videos. Researchers have proposed a number of approaches that deal with the problem of vocabulary mismatch. One of the main approaches is using a thesaurus to expand user query using synonymous words in order to retrieve more relevant items.

A thesaurus is a database of terms and their related ones. The main purpose of the thesaurus is to control language vocabulary usages. In Arabic, a thesaurus is normally called "مكنز" (plural: "مكانز") [1]. There is a number of Arabic thesauruses that have been created manually. Some of these thesauruses are available in electronic format. Building and maintaining thesauruses manually is an expensive and time consuming job and affected by the background and experience of the people who build them. Hence, there is a high demand for constructing thesauruses automatically.

In this work, we used a large document collection to automatically build an Arabic thesaurus. The thesaurus was built as a similarity thesaurus using two techniques and was evaluated by some experts. The paper is organized as follows. In Section II, we discuss different approaches for constructing thesaurus automatically. Then, we describe the similarity thesaurus approach and explain its construction and usages. In Section IV, we propose an improvement to the query expansion approach using similarity thesaurus. Section V describes our experiment and evaluation results. Finally, we present our conclusions and future work.

II. RELATED WORK

In the literature, there is a number of different proposed approaches for constructing thesaurus automatically. Research in this field started in the early sixties. During three decades, the focus was almost on co-occurrence analysis that is based on the idea that related terms co-occur frequently in the document collection. Until the beginnings of the 1990s, the approaches for constructing thesaurus automatically had not been very successful [2]. Later, several valuable approaches have been proposed in the literature.

Crouch and Yang [3] conducted an experiment on generating thesaurus automatically based on the discrimination value model and on document clustering. They mentioned that clustering of low frequent terms is not appropriate in small collections. For that, they clustered the documents of the collection in order to generate thesaurus classes from the low frequency terms contained in the document clusters. The clustering algorithm that was used is the complete-link clustering algorithm. The experiment was applied on various document collections and the result showed some improvement in retrieval performance.

Qiu and Frei [4] presented a probabilistic query expansion model based on an automatically constructed similarity thesaurus. Their approach expands queries based on similarity to their concepts rather than similarity to the individual terms in the query. To the best of our knowledge, this approach was the first approach to expand queries using their concepts. Qiu and Frei ran their experiment on several test collections. The results showed a notable improvement on the retrieval effectiveness. Zazo et al. [5] also used the same approach for constructing a Spanish similarity thesaurus. Qiu and Frei approach will be described in detail in the next section.

Grefenstette [6] proposed an alternative approach to co-occurrence analysis which is based on syntactic analysis. The basic idea of his approach is that words found in the same contexts tend to be semantically related. Grefenstette showed that his approach is better than the co-occurrence analysis approach. Similarly, Gauch and Wang [7] proposed a linguistics-based analysis technique to automatically discovering similarities. For each word, they constructed a context vector that summarizes information about terms occurrences around the word. A similarity matrix is then created by calculating terms similarities based on their context vectors. Gauch and Wang reported an improvement of the retrieval performance using their approach.

Chen and Lynch [8] carried out an experiment on applying algorithmic approach to the automatic generation of thesaurus based on statistical correlation analysis. The constructed thesaurus is captured in a semantic network where nodes represent concepts and weighted links represent their relevance strengths. The results of their approach were encouraging. Tsen [9] used a similar

approach for automatic thesaurus construction using a Chinese document collection.

Jing and Croft [10] conducted an experiment using text analysis feature recognition. Their approach determines associations between phrases and terms by viewing a text document as structured object. Similarly, Schutze and Pedersen [11] presented a method for automatically constructing a thesaurus based on lexical co-occurrence. They argued that terms that often occur within some distance of each other are most likely to be related than if they just occur in the same document. Both studies showed that these types of approaches were able to improve the retrieval performance.

Park et al. [12] used Bayesian network to construct thesaurus automatically. They mentioned that purely statistical approaches are unreliable for deciding the relationships of low frequent terms because of the data sparseness problem. Their experiment results showed retrieval effectiveness in comparison with the results of co-occurrence analysis.

III. SIMILARITY THESAURUS

A. Overview

The similarity thesaurus is a matrix that represents terms similarities. Each term is initially represented by a vector that determines its relation with each document in the collection. The matrix is then generated through calculating similarities between terms vectors.

Query expansion using similarity thesaurus is analogous to the translation from a language to another [4]. Obviously, looking for the translation of every single word in a sentence in the dictionary will not probably give an accurate translation for the whole sentence. Instead, a professional translator can express his/ her understanding by selecting the most suitable meanings. Using similarity thesaurus, a query is expanded by selecting thesaurus terms that are similar to the concept of the query as a whole rather than to individual terms in the query. In Arabic, there are many words that have different meanings based on the contexts. For example, Fig. 1 shows three different contexts of the word "قرص".

B. Construction Approach

In the similarity thesaurus construction approach as described in [4], the normal indexing roles of documents and terms are interchanged. In most of the other approaches, documents are normally indexed by terms while in this approach terms are indexed by documents. The justification of this is that the probability of a term representing a document is not identical to the probability of a document representing a term.

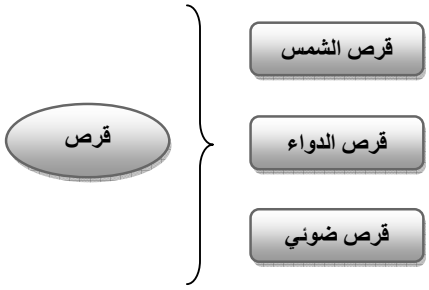

Figure 1. Different contexts of the word "قرص"

In similarity thesaurus, each term t_i is represented by a vector $\vec{t}_i$ in the document vector space, $\vec{t}_i = (w_{i1}, w_{i2}, \ldots, w_{in})$. The weight w_{ij} represents the relation between the term t_i and the document d_j. The weight w_{ij} is computed using (1).

$$w_{ij} = \frac{(0.5+0.5\frac{f_{ij}}{mf(t_i)})itf_j}{\sqrt{\sum_{l=1}^{n}(0.5+0.5\frac{f_{il}}{mf(t_i)})^2 itf_j^{\,2}}} \quad (1)$$

Where, n is the number of documents and f_{ij} is the occurrence frequency of term t_i in document d_j while mf is the maximum frequency of the term t_i in all documents. The itf_j is the inverse term frequency for document d_j and calculated using (2).

$$itf_j = \log\frac{k}{dt_j} \quad (2)$$

Where, k is the number of terms and dt_j is the number of distinct terms in the document d_j. Using inverse term frequency itf indicates that a short document plays more important role than a long one. More precisely, if two terms co-occur in a short document, the probability that the two terms are similar is greater than if they co-occur in a long document.

After that, similarity between any two terms can be calculated using the scalar vector product as described in (3).

$$sim(t_i, t_j) = \vec{t}_i * \vec{t}_j = \sum_{l=1}^{n} w_{il} * w_{jl} \quad (3)$$

Finally, the similarity thesaurus is constructed by computing the similarities for each pair of the collection terms. The constructed thesaurus is a symmetric matrix with values ranging from zero to one.

C. Query Expansion

Using similarity thesaurus, a query is expanded by looking for terms that have high relevance with the entire query terms. With the assumption that query terms have equal weights, the similarity between the query $\boldsymbol{q}$ and any term $\boldsymbol{t}$ is computed as the sum of the similarities values between each query term and $\boldsymbol{t}$ as shown in (4).

$$\text{SIM\_QT}(\boldsymbol{q}, \boldsymbol{t}) = \sum_{t_i \in q} sim(t_i, t) \quad (4)$$

The similarity values between all terms are the entries in the similarity thesaurus which already have been computed. Thus, as a response to any query, the results can be ranked in descending order according to their SIM_QT values.

IV. PROPOSED IMPROVEMENT

As we have seen, query expansion using similarity thesaurus is computed using the sum of similarity values between the query terms and each indexed term. The terms with the highest similarity values will be selected to be the most related terms. Based on our experience, we believe that this method (which we refer to as the "SUM" method) is appropriate when the similarity values between the query terms and the indexed term are consistent within

the same range. But, the question is what about if the similarity values are inconsistent. For example, the case of having one of the query terms with a high similarity value with the term ***t*** and the rest of the terms have very low similarity values. In this case, the differences between the values will not be reflected on the total sum. Thus, as a response to a query, the term ***t*** may be considered to be very related to that query even though it is only very similar to one of the terms in the query.

We believe that one of the factors that needed to be considered in query expansion is the dispersion between the similarity values. From a statistical point of view, the standard deviation can be used to indicate how the terms are distributed. However, the total similarity value should remain as the main factor. For that, after studying this problem carefully, we came up with a method that depends on the total similarity values and their standard deviation. Instead of using the sum of the similarity values, we use the mean of the values subtracted by the standard error of the mean as illustrated in (5).

$$\text{SIM\_QT}(\boldsymbol{q}, \boldsymbol{t}) = MEAN_{t_i \in q}(sim(t_i, t)) - SE \qquad (5)$$

Where, *SE* is the standard error of the mean. The standard error of the mean is a measure of data dispersion [13] and can be calculated using (6).

$$\text{SE} = \alpha / \sqrt{n} \qquad (6)$$

Where, α is the standard deviation and *n* is the number of values. We will refer to the proposed method as the "MEAN" method.

V. The Experiment

A. Overview

In order to evaluate our proposed method, we used the France Press Agency news as the document collection. This document collection can be found in the Arabic Gigaword corpus (Third Edition) from the Linguistic Data Consortium (LDC) [14]. The selected data was the news of years 2004, 2005 and 2006. We reprocessed the corpus to make each news story as a separate document since the corpus news were classified by months for each year. Also, we only used the text bodies of the news without the headings. The characteristics of the used collection are given in Table I.

After examining the high frequency terms in the collection, we had chosen 150 stop words. The collection characteristics, shown in Table I, were computed after eliminating these stop words.

TABLE I.
CHARACTERISTICS OF THE EXPERIMENT COLLECTION

Number of Documents	208,596
Number of Terms	435,846
Number of Terms Occurrences	30,415,222
Number of Processed Terms	248,311
Average Number of Words per Document	69.78

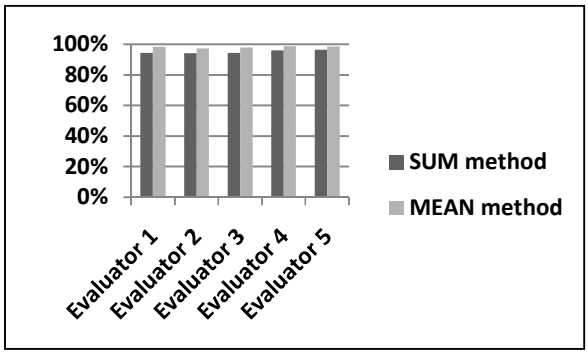

Figure 2. Evaluation results

Also, we only included terms that have occurred in more than one document. Terms that have only occurred in a single document were ignored. A total of 248,311 distinct terms were processed to generate our Arabic similarity thesaurus.

B. Evaluation

The accuracy of the thesaurus is the main issue that needs to be examined since the accuracy will dramatically affect the applications that will use that thesaurus. For that, we looked for an evaluation methodology that assesses the relevance strength of the produced thesaurus terms.

We have selected twenty common topics to be used for the evaluation process. The selected topics belong to five different domains and have been covered in the document collection. For each topic, the top ten related terms were presented to five expert evaluators. The evaluators are editorial managers from two press agencies and three newspapers. The objective of the evaluation was to assess the relevance strength of the terms that were produced by the constructed thesaurus. The evaluation process was applied for both the "SUM" and "MEAN" methods. Appendix A shows the topics and their related terms.

Each evaluator was asked to study these twenty topics carefully and then specify if the produced terms are relevant to the topics or not. We used three levels of relevance: "relevant", "somewhat relevant", and "irrelevant". *Relevant* terms take the score one and *somewhat relevant* terms take 0.5 while *irrelevant* terms take zero. After that, the evaluation data were analyzed to determine the total relevance strength of the tested query expansion methods.

The evaluation results show that the relevance strength of the standard "SUM" method was 95.0% while the relevance strength of the proposed "MEAN" method was 98.1%. Cleary, the "MEAN" method was more accurate than the "SUM" method. The evaluation results are shown in Fig. 2.

It has to be mentioned that, for neutrality, the produced terms using the "SUM" and "MEAN" methods were combined in a single list during the expert evaluation process and were ordered alphabetically.

C. Discussion

As shown in the evaluation results, all evaluators decided that the "Mean" method is more accurate than the "SUM" method. In fact, we believe that the main reason that makes the "MEAN" method a better method is its ability to detect and exclude outliers. An outlier is a value

TABLE II.
THESAURUS RESULTS USING "SUM" METHOD

Most Related Terms	Value
الفرنسي	0.814
الاليزيه	0.630
فرنسا	0.556
الرئيس	0.503
سترو	0.482

that is considerably dissimilar or inconsistent with the majority of the data [15]. Thus, an important factor that needs to be taken into account is that a given term should have a high similarity value with each individual term in the query in order to be considered related.

Query expansion using similarity thesaurus considers the concept of the query as a whole. However, adding a single term to the query may completely change the concept of the query. For that, the candidate related term should have consistent similarities with all of the query terms. For example, as a response to a query about the former French president "جاك شيراك", the five most related terms produced using the "SUM" method are presented in Table II while the results of the "MEAN" method are presented in Table III. The results show that the four most related terms are the same in both methods. However, the fifth one that was produced by the "SUM" method is the term "سترو" which is the second name of the former British foreign minister "جاك سترو". The term "سترو" was produced because it has a high similarity value with the term "جاك" regardless of its very low relatedness with the term "شيراك". As shown in Fig. 3, all produced terms except for "سترو" have consistent similarity values with the query terms.

On the other hand, the "MEAN" method considers the term "سترو" as an outlier. Thus, the term was excluded. The "MEAN" method detects the outliers because it takes the standard deviation into account through applying the standard error of the mean. A cursory look at Fig. 3, would indicate that the standard deviation of the similarity values for the term "سترو" is very high in comparison with the rest of the terms. Clearly, values consistency inversely proportional to the standard deviation.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented our work for implementing a tool that generates an Arabic similarity thesaurus automatically from a text corpus.

TABLE III.
THESAURUS RESULTS USING "MEAN" METHOD

Most Related Terms	Value
الفرنسي	0.383
الاليزيه	0.283
فرنسا	0.266
الرئيس	0.241
باريس	0.214

Figure 3. Terms similarities example

The similarity thesaurus method was chosen because query expansion using similarity thesaurus has the advantage of considering the concept of the entire query rather than the individual query terms. We have described the similarity thesaurus construction approach and explain how it can be used for query expansion. Furthermore, we have proposed our improvement in the query expansion mechanism using similarity thesaurus. In order to evaluate the proposed improvement, we have run an experiment using the Arabic Gigaword corpus. The goal of the experiment was to determine the relevance strength of the similarity thesaurus's produced terms at query expansion. The evaluation results showed that the standard "SUM" method achieved 95.0% strength of relevance while our proposed "MEAN" method achieved 98.1%. We conclude that the "MEAN" method is more accurate mainly because it can detect and exclude the outliers.

For future work, we plan to investigate more about the following research issues:

- Applying word stemming before constructing the Arabic similarity thesaurus is a subject that needs to be researched carefully. This is because Arabic language is rich in its morphology. Moreover, query expansion using similarity thesaurus deals with concepts that can be affected positively or negatively by stemming.
- Considering collocations is another issue. Collocations are sequence of words that often occur together. Producing collocations instead of terms at query expansion is an interesting research topic. Actually, this subject needs to be researched from two perspectives. First, specifying a mechanism that makes similarity thesaurus produce collocations. Second, studying the effect of the produced collections in the retrieval performance.
- Part of speech tagging or word-category disambiguation is the process of assigning words to their grammatical category. One of the research ideas is to construct a specific similarity thesaurus using a single word-category. For example, construing a similarity thesaurus for nouns only.
- Question answering is a mechanism for answering questions automatically. We believe that using similarity thesaurus in question answering is also an open area for research.

All of these ideas are very interesting ideas and we shall start work on some of them in the near feature.

REFERENCES

[1] Abdulhadi, M., "المكانز كأدوات للتكشيف واسترجاع المعلومات", Ghareeb library, Egypt, 1989.

[2] Baeza-Yates, R. and Ribeiro-Neto, B., "Modern Information Retrieval", ACM/ Addison-Wesley, New York, United States, 1999.

[3] Crouch, C. and Yang, B., Experiments in automatic statistical thesaurus construction, Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 1992.

[4] Qiu, Y. and Frei, H., Concept Based Query Expansion, Proceedings of the Sixteenth International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, United States, 1993.

[5] Zazo, A. et al, Reformulation of queries using similarity thesauri, Information Processing and Management, Volume 41, Issue 5, 2005, pp. 1163 – 1173.

[6] Grefenstette G., Use of Syntactic Context to Produce Term Association Lists for Text Retrieval, Proceedings of the Fifteenth International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 1992.

[7] Gauch, S. and Wang, J., A corpus analysis approach for automatic query expansion, Proceedings of the sixth international conference on Information and knowledge management, Nevada, United States, 1997.

[8] Chen, H. and Lynch, K., Automatic Construction of Networks of Concepts Characterizing Document Databases, IEEE Transactions on Systems, Man and Cybernetics, Volume 22, Issue 5, 1992, pp. 885-902.

[9] Tseng, Y., Automatic Thesaurus Generation for Chinese Documents, Journal of the American Society for Information Science and Technology, Volume 53, Issue 13, 2002, pp. 1130-1138.

[10] Jing, Y. and Croft, W., An Association Thesaurus for Information Retrieval, Proceedings of RIAO Conference, New York, United States, 1994.

[11] Schujtze, H. and Pedersen, O., Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, Information Processing and Management Journal, Volume 33, Issue 3, 1997, pp. 307 – 318.

[12] Fung, R., Favero, B., Applying Bayesian Networks to Information Retrieval, Communications of the ACM, Volume 38, Issue 3, 1995, pp. 42-48.

[13] Graff, D., "Arabic Gigaword Third Edition", Linguistic Data Consortium, Philadelphia, United States, 2007.

[14] Good, P. and Hardin, J., "common errors in statistics", John Wiley & Sons, New Jersey, United States, 2006.

[15] Ramaswamy, S. et al, Efficient Algorithms for Mining Outliers from Large Data Sets, ACM SIGMOD international conference on Management of data, Texas, United States, 2000.

Appendix A: Experiment's Results

Topic	Method	Top ten related terms (ordered from left to right)
مكة المكرمة	"SUM"	الحجيج - فريضة - التدافع - والمدينة - الحرام - المنورة - الحجاج - حاجا - الجمرات - الحج
	"MEAN"	فريضة - الحجيج - التدافع - والمدينة - الحرام - المنورة - الحجاج - حاجا - الجمرات - الحج
المملكة العربية السعودية	"SUM"	سعودي - الدول - السعوديين - الملك - عبدالله - عبدالعزيز - الامير - بن - السعودي - الرياض
	"MEAN"	الانباء - سعود - العاهل - الامير - الملك - عبدالله - عبدالعزيز - الرياض - السعودي - بن
كوريا الشمالية	"SUM"	النووية - لكوريا - واليابان - وكوريا - الكوري - السداسية - الجنوبية - الكورية - يانغ - بيونغ
	"MEAN"	نووية - تجربة - النووية - لكوريا - وكوريا - الكوري - السداسية - الكورية - يانغ - بيونغ
الولايات المتحدة الاميركية	"SUM"	مجلس - الامن - الامم - جورج - العراق - الرئيس - بوش - الخارجية - واشنطن - الاميركي
	"MEAN"	الاوسط - الحكومة - الامن - الرئيس - جورج - العراق - بوش - الخارجية - واشنطن - الاميركي
الملك عبدالله بن عبدالعزيز	"SUM"	الاردني - المملكة - العهد - فهد - الرياض - ولي - السعودية - الامير - السعودي - العاهل
	"MEAN"	العربية - العهد - ولي - فهد - الرياض - المملكة - السعودية - الامير - العاهل - السعودي
الشيخ احمد ياسين	"SUM"	المصري - بن - الرنتيسي - الاسلامية - محمد - الروحي - حماس - رئيس - شرم - مؤسس
	"MEAN"	الروحي - محمد - رئيس - اسرائيل - اغتيال - حركة - المقاومة - الاسلامية - مؤسس - حماس
صدام حسين	"SUM"	محاكمة - الرئيس - بغداد - السابق - العراقية - الدجيل - نظام - العراق - العراقي - المخلوع
	"MEAN"	محاكمة - الرئيس - بغداد - الدجيل - السابق - نظام - العراقية - العراق - المخلوع - العراقي
محمود احمدي نجاد	"SUM"	الفلسطينية - تخصيب - اليورانيوم - النووي - الايرانية - الرئيس - عباس - طهران - ايران - الايراني
	"MEAN"	نووي - النووية - تخصيب - اليورانيوم - النووي - الايرانية - الرئيس - طهران - ايران - الايراني
زلزال تسونامي	"SUM"	المحيط - الزلازل - ضرب - درجات - سومطرة - امواج - مقياس - المد - ريشتر - الزلزال
	"MEAN"	درجات – مقياس - المحيط - الاندونيسية - البحري - اندونيسيا - امواج - الزلزال - سومطرة - المد
اغتيال رفيق الحريري	"SUM"	التحقيق - الاسبق - لسوريا - سوريا - لبنان - بيروت - ميليس - اللبنانية - ديتليف - اللبناني
	"MEAN"	الاسبق - لسوريا - التحقيق - بيروت - لبنان - سوريا - اللبنانية - ميليس - ديتليف - اللبناني
التمرد في دارفور	"SUM"	البشير - تشاد - الجنجويد - والمساواة - ابوجا - الافريقي - السوداني - الخرطوم - السودانية - السودان
	"MEAN"	الافريقي - تحرير - حركتي - السوداني - ابوجا - الخرطوم - المتمردين - والمساواة - السودانية - السودان
البرنامج النووي الايراني	"SUM"	برنامجها - الايرانية - تخصيب - للطاقة - النووية - الملف - الذرية - طهران - اليورانيوم - ايران
	"MEAN"	الايرانية - برنامجها - الملف - تخصيب - للطاقة - طهران - النووية - الذرية - اليورانيوم - ايران
جامعة الدول العربية	"SUM"	الرئيس - رئيس - مجلس - وزير - العام - دول - عمرو - الخارجية - الجامعة - المتحدة
	"MEAN"	وزير - مجلس - الولايات - عمرو - الخارجية - الرئيس - رئيس - العام - المتحدة - الجامعة
الامم المتحدة	"SUM"	العام - الخارجية - الدولي - الدولية - الامن - كوفي - مجلس - انان - للامم - الولايات
	"MEAN"	لبنان - العام - الامن - الدولية - الامين - الدولي - مجلس - كوفي - انان - للامم
منظمة الدول المصدرة للنفط	"SUM"	للبرميل - الخام - يوميا - اسعار - الاسعار - الانتاج - النفط - برميل - انتاجها - اوبك
	"MEAN"	مليون – الطاقة - يوميا - اسعار - الاسعار - الانتاج - النفط - برميل - انتاجها - اوبك
الوكالة الدولية للطاقة الذرية	"SUM"	الملف - طهران - الايراني - تخصيب - للوكالة - ايران - البرادعي - النووية - اليورانيوم - النووي
	"MEAN"	الملف - طهران - تخصيب - الايراني - للوكالة - البرادعي - النووية - ايران - اليورانيوم - النووي
حركة المقاومة الاسلامية	"SUM"	فتح - محمود - الحركة - الاسرائيلي - لحركة - اسرائيل - غزة - الفلسطيني - الفلسطينية - حماس
	"MEAN"	السلطة - الحركة - الاسرائيلي - محمود - لحركة - غزة - الفلسطيني - اسرائيل - الفلسطينية - حماس
حزب كاديما	"SUM"	بيريتس - لبنان - الحزب - الاسرائيلي - لحزب - اسرائيل - الليكود - الله - ايهود - اولمرت
	"MEAN"	اليميني - عمير - بيريتس - مقعدا - الكنيست - التشريعية - حزبه - ايهود - اولمرت - الليكود
حزب الله	"SUM"	الاسرائيلية - وحزب - رئيس - الاسرائيلي - لحزب - اللبنانية - اسرائيل - الشيعي - اللبناني - لبنان
	"MEAN"	الاسرائيلية - وحزب - رئيس - لحزب - الاسرائيلي - اسرائيل - اللبنانية - الشيعي - اللبناني - لبنان
تنظيم القاعدة	"SUM"	الانترنت - اعتداءات - زعيم - بن - بلاد - الرافدين - مصعب - الزرقاوي - اسامة - لادن
	"MEAN"	الانترنت - الظواهري - بن - زعيم - بلاد - الرافدين - مصعب - الزرقاوي - اسامة - لادن