Clustering Human Association

Raz Tamir, School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel imr@netvision.net.il

Abstract

We combined a novel associations strength ranking algorithm and an unsupervised Self Organizing Maps technique to cluster free associations. We tested the algorithm on 171 seed terms and showed a very good clustering performance. The algorithm suggests a linkage between the two biggest known databases - the human mind and the Internet. The cluster labels can be used as highly informative aspects of the seed terms used for query expansion and off-line data retrieval algorithms.

Key Words

Associative Memory, Associations Generation, Clustering, Classification, Data Mining, Human Perception and Communication, Statistical Methods in AI, Web Intelligence.

1. Introduction

Associations are links between knowledge units inside human memory. Knowledge units can be concepts, images, smells etc. Experiments proved associations to be an efficient tool for transferring ideas between people [Murakami, 2001]. Thus, it is very interesting to utilize the associative mechanism as an alternative link or shortcut between knowledge units, both in human mind and in computer databases. One of system that could benefit enormously from this new linking mechanism is the Internet.

The Internet brings a vast collection of small documents dealing with almost every subject known to man. The enormity of data makes it almost impossible to retrieve relevant information without the usage of search engines. Unfortunately, even search engines can give but a narrow glimpse into web content since they are limited both by format (most of search results produce lists of pages with short description of each) and by context (the need to present relevant information out of practically infinite database where search terms is usually not well defined). While running a search people usually

look at the few top matching pages and are not exposed to lower scored www pages. Thus, the issue of information retrieval must be addressed both the format and context directions.

1.1. Retrieving relevant data sets

The only database people use, which is bigger than the Internet is their own brain. Researches showed that associations are among the most powerful tools employed while retrieving information from the brain. In previous research [Tamir, 2003] we presented an algorithm for association generation and scoring. The algorithm was based on a novel Confidence Gain ("CG") measure, and was able to extract association from any database including the Internet. We showed that the CG measure extracts terms that are equivalent to terms extracted through free associations processes.

1.2 Unsupervised clustering methods

Two main approaches exist for unsupervised clustering: partitioning and hierarchical clustering. Hierarchical clustering creates nested representation of data. The hierarchies are formed by merging and splitting clusters, according to a similarity metric. A standard hierarchical clustering algorithm is AGENS (Agglomerative Nesting). It starts with n clusters containing one element each. Then one basic step is repeated until converging - merging of the two most similar clusters. The complexity of AGNES is $o(n^2\log(n))$. An opposite approach is presented in DIANA algorithm where a single cluster containing all elements is divided various times until a tree of clusters is formed. Both algorithms have variants where the similarity measure is different.

The most used unsupervised partitioning clustering algorithm is the K-Mean. The algorithm assumes K initial clusters seeds (or cluster centers). Each element is assigned to a cluster. Then the cluster center is re-computed as the average of its elements. Element assignment and cluster re-computation are repeated until converging.

The K-Mean algorithm is an efficient one, having a linear complexity. It has many variants, such as "Bisecting K-means" and "K-medoids" [Pantel, 2003]. Several algorithms use a hybrid version of the hierarchical and partitioning clustering.

There are other approaches to clustering. For example in CBC (Clustering By Committee) [Pantel, 2002] where cluster centroids are set by a subset of the cluster

members (called committee) who also determines which other elements belongs to the cluster. A different yet interesting example is "Gravitational Clustering" [Gomez, 2003].
In the last two decades there was a great development in the neural-networks field. One of the pioneering algorithms known as Self Organizing Map (SOM) [Kohonan, 1982] has proven very powerful in clustering high dimensional data.
The SOM algorithm is used to form a condensed description of a data set consisting of multivariate observations. SOM manipulates a set of model vectors attached to nodes on a 2D map and trains it while trying to follow the distribution of the input data. Neiboring positions in the 2D map are kept close thus having a close conceptual interpretation [Kashi, 1998]. One of SOM second-generation algorithms, WEBSOM [Honkela, 1996] is specifically used for clustering and navigation through big collections of documents. WEBSOM uses two layers where the lower is a SOM of “term categories” and the upper is the clustered documents collection. The “term categories” serve as document topics. Zooming into the map and navigation through it enable intuitive browsing of the user until a relevant document is retrieved.
The basic problem of such clustering is the tendency to treat similar but not identical words as being totally different. This phenomena result in large dimension vectors which contain many null elements. Beside of the growing computational complexity, common similarity measures tend to be less precise. In order to address this issue it is possible to use anthologies as background knowledge. A considerable improvement in clustering results can be achieved by using some ontological strategies [Honto, 2003]. WordNet, an on-line lexical reference system [Miller, 1993] can also be used as a disambiguation aid and detect similarities between terms. In the current research we used association measures to calculate the similarity between pairs of terms. The association measures perform the disambiguation task by assigning high association strength to related but not identical words.

1.3 Current research contribution

In the current research we combined association-scoring algorithm based on CG and SOM in order to form a self-clustering map of human associations.

The association clusters represent different senses; such as exist in free association process. We used the "University of South Florida Homograph Norms" database to measure the clustering success. Homographs are words with identical spelling and two or more distinct meanings. The database consists of an experiment results. During the experiment 320 words were selected from Roget's International Thesaurus (1962). 46 students were instructed to write the first word they thought of after being presented with the list. The associations were then manually collected into separate categories. The collection of manually categorized free associations made the FAN ideal for human association clustering.

2. Clustering associations using SOM and CG algorithms

2.1 SOM clustering algorithm

One of the usages of SOM is creating "Contextual Maps" or "Semantic Maps" of textual databases such as newsgroups. The basic SOM algorithm places a set of reference vectors, called model vectors, into a data space [Lagos, 2000] . The algorithm ends when the updated model vectors approximate a given data set. The data space is usually a rectangular or hexagonal shaped two-dimensional surface. The data set members are presented to the algorithm in random order, several times. In each iteration the best matching model (winner) for the current member is searched. Subsequently, the winner model and its neighbors are updated. The common updating formula is of the form

[2.1] $$m_i(p,t+1) = m_i(p,t) + [x_i - m_i(p,t)] \cdot r(p,t)$$

Where $m_i(p,t)$ is the value at the i^{th} element of a model vector located at position p during iteration t, x_i is the value of the element at the i^{th} position of the data-set member vector, and r represents the learning rate of the SOM. r is dependent on the iteration and the position of the adjusting node.

When used for text documents clustering, each word (after removal of ASCII drawings and also words that appear very often or rarely) is given a unique random unit-norm n-dimensional vector. Let a word be indexed by k, and represented by the unique sample vector r_k. All occurrences of the word are then scanned. The location of the word k is marked by j(k). An "average context vector" of the word k is formed as

[2.2]
$$x_k = \begin{bmatrix} E\{r_{j(k)-1}\} \\ e \cdot r_{j(k)-1} \\ E\{r_{j(k)+1}\} \end{bmatrix} \in \Re^{3n}$$

Here E is the average over all j(k), and e is a scaling parameter.

The words tend to be clustered into "word categories". Such a system is commonly referred to as WEBSOM [Kohonen, 2001.]. Several attempts were made to reduce the dimensionality of document vectors, such as Latent Semantic Indexing (LSI) [Kohonen, 2000] or Randomly Projected Histograms.

2.2. Scoring associations using Confidence Gain measure

Given a stimulus word X and another word Y (called *associative response*), two well-defined measures describe the degree of support (Supp) and confidence (Conf) of the association between them:

[2.3]
$$\mathrm{Supp}(X \Rightarrow Y) = \mathrm{Supp}(X \cup Y) = \frac{\|X \,\&\, Y\|}{w}$$
$$\mathrm{Conf}(X \Rightarrow Y) = \frac{\mathrm{Supp}(X \cup Y)}{\mathrm{Supp}(X)}$$

In this formula, w is the total number of scanned web pages, ||X&Y|| is the number of web pages containing both X and Y, and Supp(X) is the fraction of pages in the world wide web that contain word X. For a fixed stimulus, Conf is a function of all possible responses Y.

The average confidence of response Y, $\overline{\mathrm{Conf}(Y)}$, is calculated as follows:

[2.4]
$$\overline{\mathrm{Conf}\ (Y)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{Supp}(I_i \cup Y)}{\mathrm{Supp}(I_i)}$$

Hereby, n is the number of valid instances. An instance is valid if support and confidence are above certain thresholds. Given the above definitions, the confidence gain measure (CG) is a function of both X and Y:

[2.5]
$$\mathrm{CG}(X \Rightarrow Y) = \frac{\mathrm{Conf}(X \Rightarrow Y)}{\overline{\mathrm{Conf}(Y)}} = n \cdot \frac{\dfrac{\mathrm{Supp}(X \cup Y)}{\mathrm{Supp}(X)}}{\displaystyle\sum_{i=1}^{n} \frac{\mathrm{Supp}(I_i \cup Y)}{\mathrm{Supp}(I_i)}}$$

This formula implies that the co-occurrence frequency of a stimulus/response pair is compared to an average co-occurrence frequency of other word pairs. In these other word pairs the response is always the original one and only the stimulus varies.

Our algorithm uses the API of the well-known Google search engine for retrieving relevant documents from the World Wide Web. We calculate all pair's co-occurrences throughout the web by performing multiple queries, where only the number of sites containing the pairs is retrieved, rather than the sites themselves.

2.3 Clustering associations algorithm

2.3.1 Data Preprocessing

In order to represent the relation between all terms by the CG measure, we divided all terms into all possible pairs. For each pair CG was calculated by [2.5]. The logarithm of the CG values was used for more convenient usage.

Now, we represented each term by a vector containing its relation to each of the other terms by the $\log_{10}(CG)$. All numbers grater than 1 (where CG was greater than 10) were set to 1. All numbers smaller than -1 (where CG was smaller than 0.1) were set to -1). Thus given k initial terms, we formed k vectors of k elements each. These vectors were used to train the SOM. We initialized the Self Organizing Map to a 10X10 cells. We linked a k-element vector to each cell. Each element was randomly set to values ranging from -1 to 1.

2.3.2 Training the SOM

The training process consisted of 600 iterations. As shown (see section 2.1) we selected the formula of the form

[2.1] $$m_i(p,t+1) = m_i(p,t) + [x_i - m_i(p,t)] \cdot r(p,t)$$

Where $m_i(p,t)$ is the value at the i^{th} position in a node located at position p during iteration t, x_i is the value of the current term at the i^{th} position and r represents the learning rate of the SOM.

The two dimensional map was trained according to the following procedure:

1. Go over all term vectors. For each term:

a. Find the node on the map having the smallest Euclidian distance to the term's vector.
b. Train all the nodes in the SOM using the formula [2.1]

For the first one hundred iterations the Learning Rate r was calculated as:

[2.6] $$r = 0.05 \cdot e^{\sqrt{\Delta x^2 + \Delta y^2}/10}$$

We trained the entire map nodes where $\sqrt{\Delta x^2 + \Delta y^2}$, the learning radius, was calculated as the Euclidian distance between a given node on the map and the node selected at stage 1a of the algorithm.

After 100 iterations both Learning Radius and the Learning Rate itself were decreased using:

[2.7] $$r = 0.01 \cdot e^{\sqrt{\Delta x^2 + \Delta y^2}/50}$$

At this stage we trained only neighboring cells.

2.3.3 Labeling Clusters

After 600 iterations each term has "chosen" a specific node on the map. The map contained clusters of terms. Each node that was related to more terms than any of his neighbors was selected as a cluster center. All terms in the cluster center node and the neighboring nodes were assumed to be part of the cluster. In order to choose the best label for each cluster we compared three scoring criteria:

The first was based on selecting the word having the highest CG to the source term. The underlying assumption is that the centers of different clusters represent different senses of the source term. Previous research [Tamir, Rapp 2003] showed that senses of an ambiguous word are best described by terms that, although bearing a strong association to this word, are mutually exclusive, i.e. whose association strength is as weak as possible. The exclusiveness is reached by selecting terms from different clusters.

The second criterion was selecting the closest word to the cluster center, by performing Euclidian distance calculation. This criterion selects the term having minimal distance to the cluster center.

The third criterion was based on the LableSOM approach [Rauber, 1999]. After gathering all the terms related to a cluster, we calculated the average distance between each element

in the vectors the element in the cluster center vector. The element that was, in average, closest to the element in the same position in the cluster center vector was chosen as a cluster label.

3. Results

3.1 Data preprocessing

We applied the algorithm over a free association homograph norms database [Nelson, 1980] (denoted Free Association Norms or FAN). In order to measure the clustering algorithm success we chose all the seed terms that had more than one sense, and all senses having more than 3 related associations (the latter constraint was set in order to get enough associations to create a cluster).

3.1 Creating SOM

After filtering the database we used 171 seed terms and 3504 valid associations to create separate SOM for each term. We labeled the clusters using the three discussed methods. Finally we measured the quality of clusters formation, labeling and sense disambiguation. The following figure shows the cluster formed for the term "ORGAN". The following clusters can be spotted:

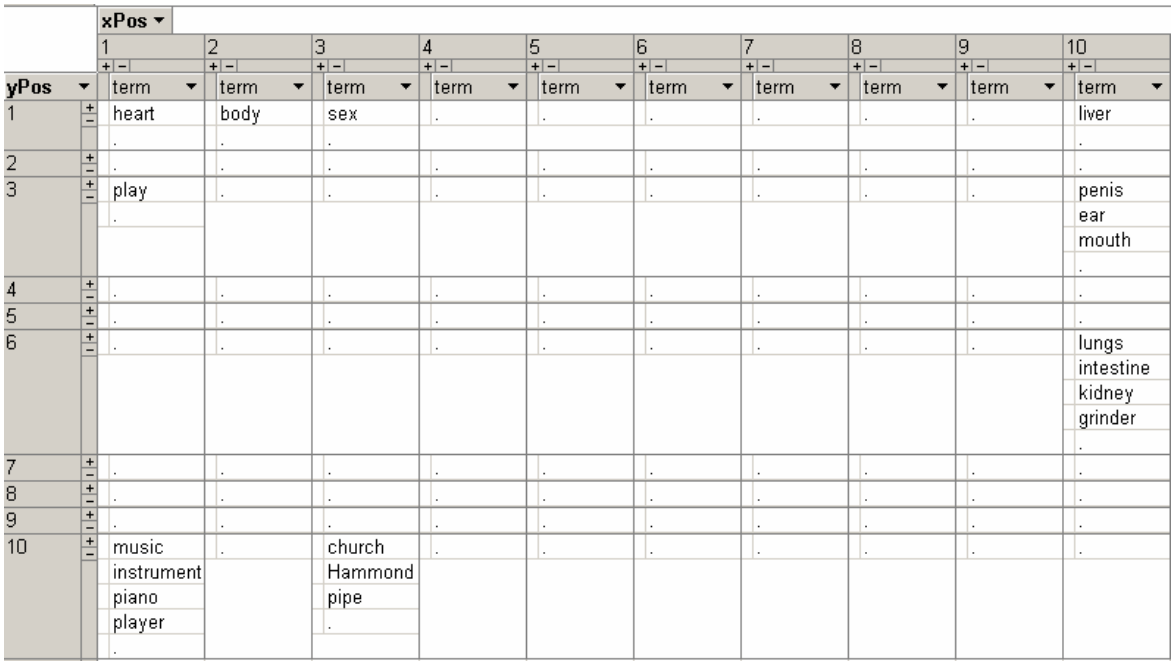
{"music","instrument","piano","player"}

{"church","hammond","pipe"}

{"penis","ear","mouth"}

{"lungs","intestine","kidney","grinder"}

{"heart","body","sex"}



yPos \ xPos	1	2	3	4	5	6	7	8	9	10
	term	term	term	term	term	term	term	term	term	term
1	heart	body	sex	.	.	.	.	.	.	liver
	.	.	.							.
2	.	.	.	.	.	.	.	.	.	.
3	play	.	.	.	.	.	.	.	.	penis
	.									ear
										mouth
										.
4	.	.	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.	lungs
										intestine
										kidney
										grinder
										.
7	.	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.	.
10	music	.	church	.	.	.	.	.	.	.
	instrument		Hammond							
	piano		pipe							
	player		.							
	.									

Figure 1 Clusters formed for the word "ORGAN"

3.3 Identifying and Labeling Clusters

In order to grade the clustering success we did the following:

1. Select minimal allowed cluster size. For an example, if minimal cluster size is 1 then the term "Play" in figure 1 is a cluster, and the figure consists of 7 clusters. On the other hand if we set the minimal cluster size to 4 than only 2 clusters exist.
2. Identify clusters centers coordinates.
3. Classify each term as a member of the closest cluster center.
4. Choose labeling method (See section 2.3.4) and for each cluster select the appropriate term (association).
5. Label each cluster after the Sense of the selected term. For example if for the cluster in position [1,10] the chosen term is "music", than the cluster will be labeled "instrument", since the association "music" is classified under "instrument" in the FAN database.

6. Classify each term after the label of the cluster it belongs to. For example the terms "instrument", "piano" and "player" will now get the label "instrument".

3.4 Grading Classification Results

Three factors are commonly used for classification assessment: Recall, Precision and Accuracy [Alvarez, 2002]. We implemented these measures using the following definitions:

By examining a single term from the FAN database (e.g. "Organ") we get n associations. Let us select a single class (e.g. "instrument"). Denote F^+ as the associations that are classified as related to "instrument" by FAN database. F^- is the group of associations that were classified as related to a different class (but still are associations of "Organ").

S^+ is the group of associations that are classified as related to "instrument" by the current SOM algorithm. Finally, S^- is the group of associations of "Organ" that were classified as related to a different class.

Now, let us define:

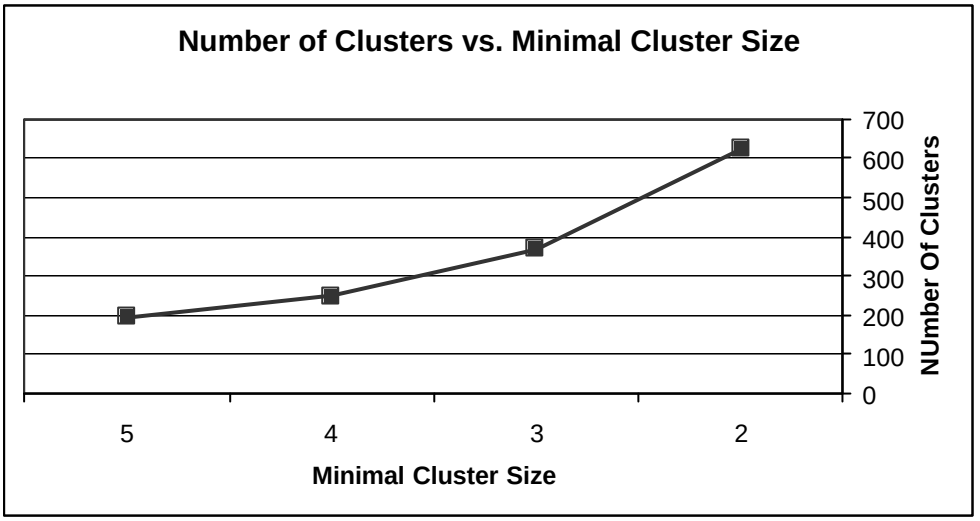
[3.1]
$$n_1 = \#\left(F^+ \cap S^+\right)$$
$$n_2 = \#\left(F^- \cap S^+\right)$$
$$n_3 = \#\left(F^+ \cap S^-\right)$$
$$n_4 = \#\left(F^- \cap S^-\right)$$

Where # means "number of items". Now we can define Precision, Recall and Accuracy as

[3.2]
$$\mathrm{Pr}\,ecision = \frac{n_1}{n_1 + n_2}$$
$$\mathrm{Re}\,call = \frac{n_1}{n_1 + n_3}$$
$$Accuracy = \frac{n_1 + n_4}{n}$$

In order to be able to compare SOM results with initial FAN database classifications we classified each cluster as the sense given by FAN database to the term we chose as label in section 2.3.4. In order to compute the overall Recall, Precision and Accuracy we averaged the results over all the clusters.

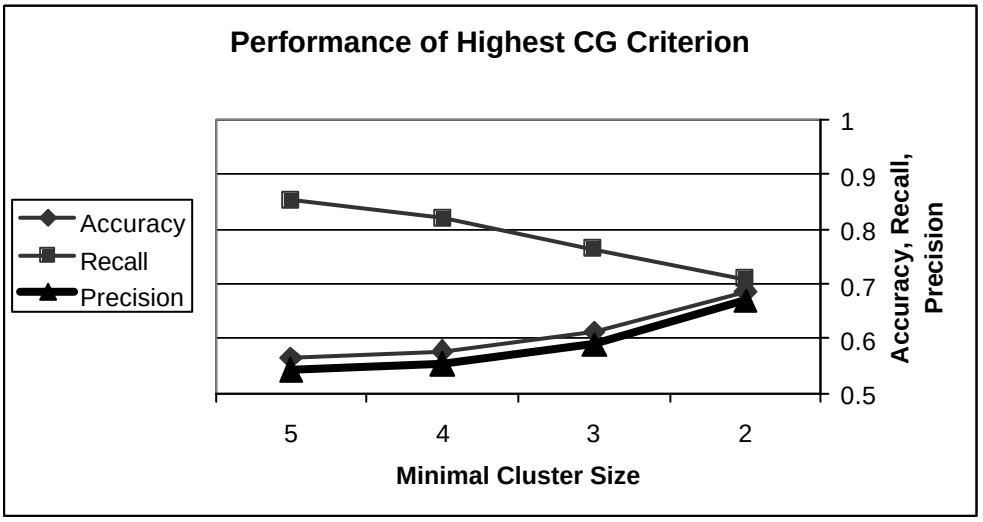
The connection between the minimal cluster size and the total number of clusters is given in the following graph.



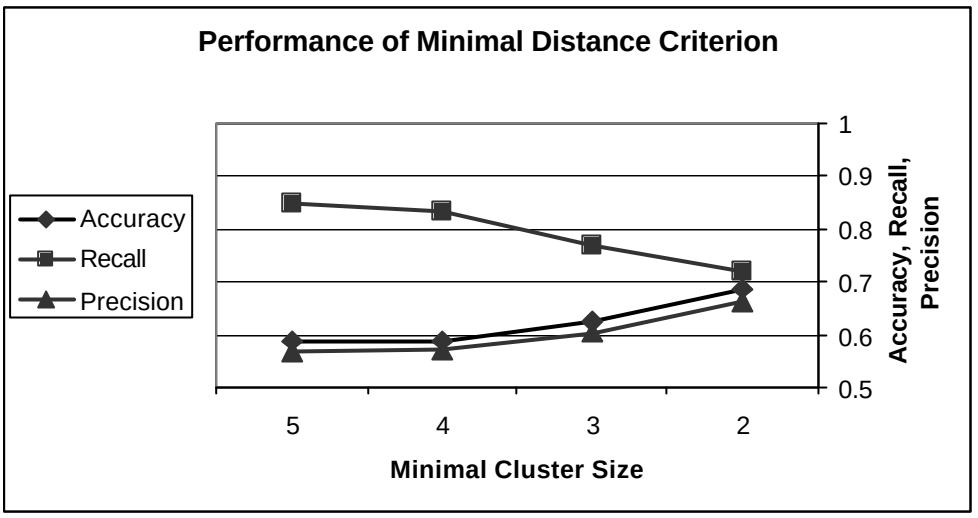
Graph 1 Number of clusters vs. minimal cluster size.

Note that in the FAN database there are 369 senses. Thus, if we allow clusters smaller than 3 terms we must have cases where two of more clusters belong to the same sense.

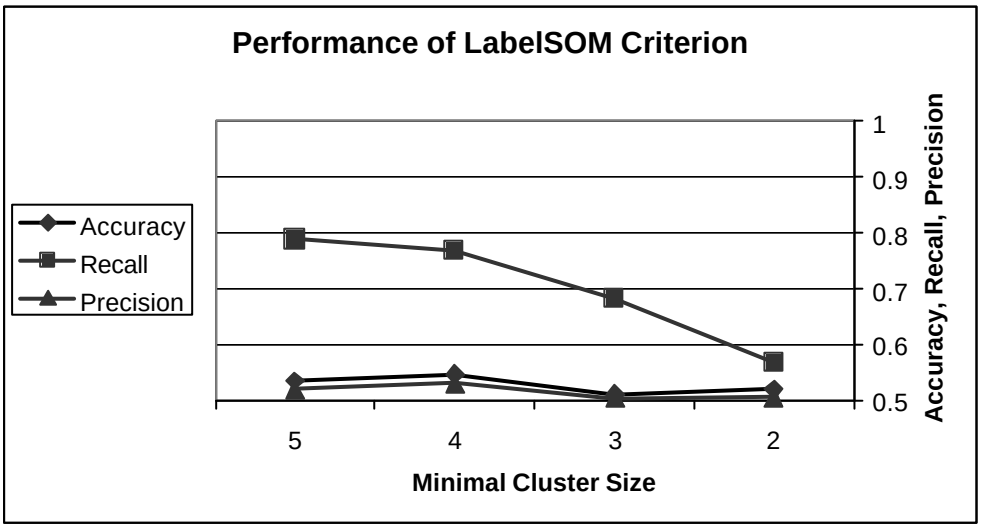
In the following graphs we calculate Recall, Precision and Accuracy for each labeling method (see section 2.3.3), and for different minimal cluster size.



Graph 2 Recall, Precision and Accuracy vs. Minimal cluster size for the "highest CG" labeling criterion.

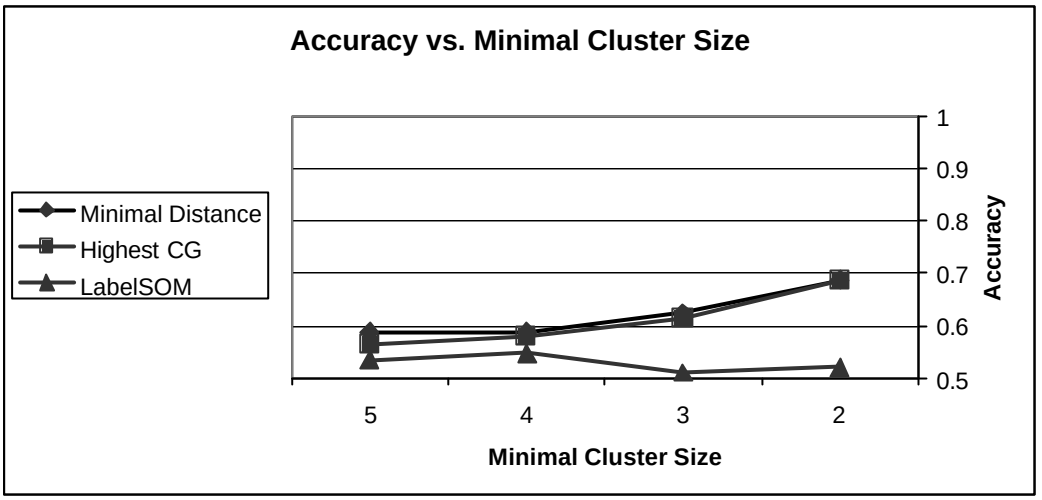


Graph 3 Recall, Precision and Accuracy vs. Minimal cluster size for the "minimal distance to cluster center" labeling criterion.

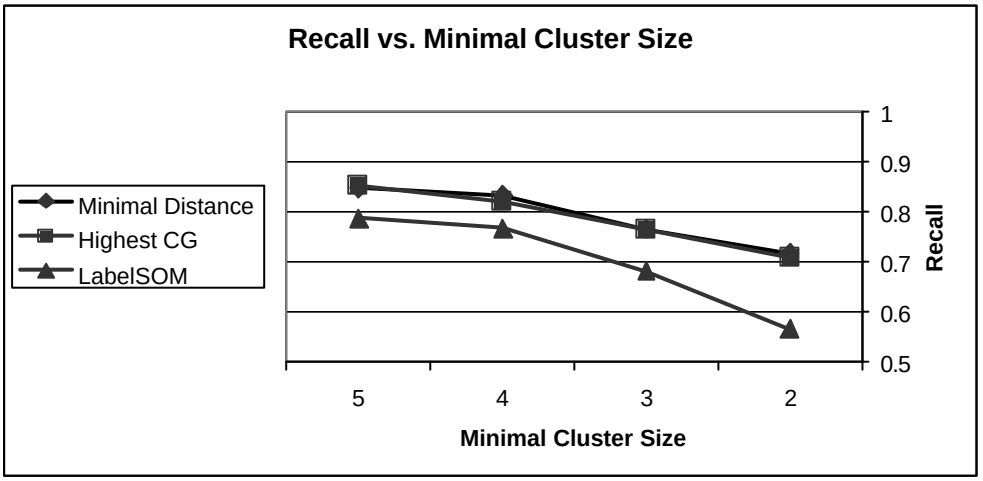


Graph 4 Recall, Precision and Accuracy vs. Minimal cluster size for the "LabelSOM" labeling criterion.

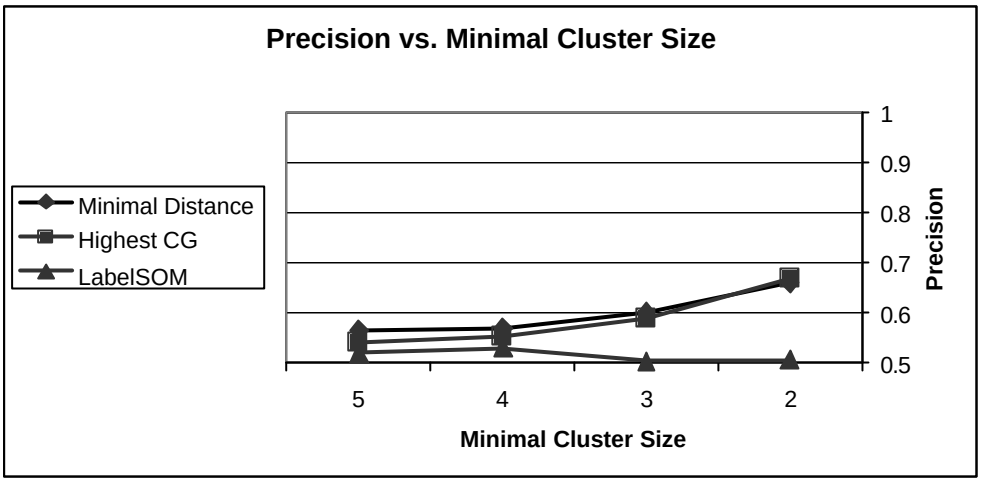
Finally, we compared the three criteria performance.



Graph 5 Accuracy vs. Minimal cluster size for the three labeling criteria.



Graph 6 Recall vs. Minimal cluster size for the three labeling criteria.



Graph 7 Precision vs. Minimal cluster size for the three labeling criteria.

3.5 Results Interpretation

3.5.1 General Performance

In order to measure the clustering success let us recall the goal of the clustering as stated in the introduction. We selected clusters labels as highly informative terms. The retrieval scenario is made of the following steps:

1. The user enters a seed term (or a seed phrase) to a search engine.
2. The highest ranked pages suggested by the search engine are parsed. All terms that appear frequently enough are inserted into a "candidate list".
3. SOM is applied to form clusters of the candidate list.
4. All clusters labels are presented back to the user as highly informative terms related to his initial seed term.

There are two basic conditions needed for the clustering and labeling algorithm to be useful to the user: (i) the suggested list of terms must be short; (ii) the suggested list should represent as many aspects of the seed term as possible. The first condition requires as few clusters as possible (since all cluster' labels are in the suggestion list). The second condition requires a high precision grade, so the cluster labels will be good representative of all cluster members. We claim that precision measure is much more important than recall for our task. Since the amount of data in the Internet is so vast, the question of how many of the relevant terms were retrieved seem less important than the question whether the retrieved terms are relevant.

The final balance between number of suggested terms and the algorithm precision is left for the user. By choosing a minimal cluster size of 2, a good classification is achieved (by selecting "minimal distance to cluster center" criterion a 68.6% accuracy, 71.9% recall and 66.1% precision are obtained). The price in this case is the large number of clusters (623). On the other hand if the user is limited by the number of clusters a choice of a minimal number of 3 terms per a cluster yields reasonable classification results (62.5% accuracy, 76.6% recall and 60.2% precision), where number of clusters is only 369 - which is the exact number of possible senses (10.5% of all terms).

3.5.2 Best Labeling Criterion

While the LabelSOM is distinctively less efficient for current usage, the "highest CG relative to stimulus" and "minimal distance to cluster center" perform similarly.

We prefer the "minimal distance to cluster center" approach for cluster labeling. Besides of its good results "minimal distance to cluster center" criterion has another important benefit over the "highest CG to stimulus" criterion. In practical situations, the stimulus is usually not given, thus it is not possible to use the "highest CG to stimulus" criterion, while the "minimal distance" criterion is not affected by the absence of known stimulus.

4. Conclusions and future work

The proposed approach combined of SOM clustering based on CG measure performs well for associations clustering. The classification performances depend on the user needs.

While comparing the current method to WEBSOM approach, we saw one distinctive difference. The WEBSOM approach looks into the immediate vicinity of the tested term, our algorithm looks for pairs appearing in the same document. This difference suggests that we can combine both approaches in order to get more data on the relations between terms. In further research we will examine the possibility to insert the "vicinity" notion into our algorithm. We think that the added information will provide knowledge of some syntactical relations between terms and enhance the clustering performance.

The biggest challenge of applying SOM for associations clustering is creating an incremental model of SOM. The need to recalculate the map to incorporate changes

(adding or excluding terms) makes it too slow to be used for real time applications. In farther research we intend to confront two main issues: performing an incremental calculation of SOM and defining a simpler and faster CG calculation. Solving these two problems will enable the use of SOM for practical on-line user true conceptual tracking.

Bibliography

1. Alvarez. S.A., *"An exact analytical relation among recall, precision, and classification accuracy in information retrieval"*, Technical Report BC-CS-2002-01, Computer Science Department, Boston College, Chestnut Hill, MA 02467 USA, June 2002. http://www.cs.bc.edu/~alvarez/APR/aprformula.pdf
2. Gomez J., Dasgupta D., Nasraoui O., *A New Gravitational Clustering Algorithm*, Proceedings of the Third SIAM, 2003 (forthcoming)
3. Honkela T., Kaski S., Lagus K., Kohonen T., *Newsgroup Exploration with the WEBSOM Method and Browsing Interface*, Technical Report, Helsinki University, Neural Networks Research Center, Espoo, Finland, 1996.
4. Hotho A., Staab S., G. Stumme G., *Ontologies Improve Text Document Clustering,*. Proceedings of the International Conference on Data Mining - ICDM-2003. IEEE Press, 2003.
5. Kaski S., et al. *"Statistical Aspects of the WEBSOM System in Organizing Document Collections"*. Computing Science and Statistics. Vol. 29. pp. 281-290, 1998.
6. Kohonen T., "*Self-Organizing Maps*", Information Sciences. Springer, third edition, 2001.
7. Kohonen T., *"Self Organized Formation of Topologically Correct Feature Maps*", Biological Cybernetics 43, 59, 1982.
8. Kohonen T., Kaski S., Lagus K., Salojarvi J., Honkela J., Paatero V., Saarela A., *"Self Organization of a Massive Document Collection"*. IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery, 11(3):574-585, 2000.
9. Kohonen T., Niklasson L., Boden M., Ziemke T.(editors), "*Self-Organization of Very Large Document Collections: State of the Art"*, Proceedings of ICANN98,

the 8th International Conference on Artificial Neural Networks, vol. 1, Springer, London, 1998, 65-74

10. Lagos K. "*Text Mining with the WEBSOM*". Acta Polytechnica Scandinavica, Mathematics and Computing Series no. 110, Espoo 2000, 54 pp. Published by the Finnish Academy of Technology. ISBN 951-666-556-X. ISSN 1456-9418. UDC 004.032.26:025.4.03:004.5, 2000.
11. Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K., *Introduction to WordNet: An Online Lexical Database*. Published on World Wide Web: ftp://clarity.princeton.edu/pub/wordnet/5papers.ps, 1993
12. Murakami H., *Cognitive Experiment on Generation and Understanding of Associative Representation*, Bulletin of Osaka City University Media Center, Vol.2, pp.3-10, March 2001.
13. Nelson D. L., McEvoy C. L., Schreiber T. A., "*The University of South Florida Word Association, Rhyme, and Word Fragment normsNorms"*, http://w3.usf.edu/FreeAssociation/Intro.html, (current February 2002).
14. Nelson, D.L., McEvoy, C.L., Walling, J.R., and Wheeler, J.W., "*The University of South Florida Homograph Norms"*, Behavior Research Methods and Instrumentation, 12, 16-37, 1980.
15. Pantel P., *Clustering by Committee*. Ph.D. Dissertation. Department of Computing Science, University of Alberta. 2003.
16. Pantel P., Lin D., *Discovering Word Senses from Text*, In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002. pp. 613-619. Edmonton, Canada, 2002.
17. Rauber A. "*LabelSOM: On the Labeling of Self-Organizing Maps"*. In Proceedings of the International Joint Conference on Neural Networks (IJCNN'99), Washington, DC, 1999.
18. Tamir R., *"Association Rules Generation and Scoring using Confidence Gain on Internet Pages*", to be published, 2002.
19. Tamir R., Rapp R. *"Mining the Web to Discover the Meaning of an Ambiguous Word*", IEEE ICDM, 2003 (forthcoming)