Detecting Sub-Topic Correspondence through Bipartite Term Clustering

Zvika MARX
The Center for Neural Computation,
The Hebrew University of Jerusalem
and
The Institute for IR and Comp. Linguistics,
Mathematics and Computer Science Dept.,
Bar-Ilan University
Ramat-Gan 52900, Israel,
Marxzv@cs.biu.ac.il

Ido DAGAN
The Institute for Info. Retrieval and Computational Linguistics,
The Department of Mathematics and Computer Science,
Bar-Ilan University
Ramat-Gan 52900, Israel,
dagan@cs.biu.ac.il

Eli SHAMIR
Institute of Computer Science
The Hebrew University of Jerusalem
Jerusalem 91904,
Israel
shamir@cs.huji.ac.il

Abstract

This paper addresses a novel task of detecting sub-topic correspondence in a pair of text fragments, enhancing common notions of text similarity. This task is addressed by coupling corresponding term subsets through bipartite clustering. The paper presents a cost-based clustering scheme and compares it with a bipartite version of the single-link method, providing illustrating results.

1. Introduction: Corresponding Entities in Text Fragments

Information technology is continuously challenged by growing demand for accurate performance in fields such as data retrieval, document classification and knowledge representation. A typical task in these areas requires well-developed capabilities of assessing similarity between text fragments (see, for example, Chapter 6 in Kowalski, 1997). Nevertheless, it is apparent that standard methods for detecting document similarity do not cover the whole range of what people perceive as similar.

Common treatment of document similarity typically aims at a unified pairwise measure expressing the extent to which documents are similar to each other. Consequently, an Internet-surfer hitting the *"what's related"* button in her browser gets a list of pages that are supposed to be similar to the currently viewed one. Here, we originally address questions situated one-step ahead: Given documents that are already known to be similar, *how* they are related to each other? How to refer a user to relevant aspects in a large collection of similar documents? One possible strategy, rooted in cognitive considerations, is to present for each pair of similar documents a detailed "map", connecting corresponding concepts, entities or sub-topics. The present article provides initial directions towards identification of such correspondences.

Consider, for example, the following two fragments taken from a pair of 1986 Reuters' news-articles:

1. LOS ANGELES, March 13 - *Computer Memories Inc.* ... agreed to **acquire** *Hemdale Film Corp.* ... That company's **owner**, *John Daly*, would then ***become*** **chief executive officer** of the combined company...
2. NEW YORK, March 25 - *Messidor Ltd* said it signed a letter of intent to **acquire** 100 pct of the outstanding shares of *Triton Beleggineng Nederland B.V.* ... If approved, the **president** of *Triton*, *Hendrik Bokma*, will be ***nominated*** as **chairman** of the combined company. ...

Both fragments deal with the intention of a certain company to acquire another company. Since the word *'acquire'* appears in both articles, keyword-based methods would interpret it as a positive evidence for evaluating the text fragments as similar to each other. More sophisticated methods (e.g. *Latent Semantic Indexing*; Deerwester et al., 1990) incorporate vector-based statistical term-similarity models that may take into account correspondence of different terms that resemble in their meaning. For example, the corresponding term pairs *'owner'* — *'president'*, and *'chief executive officer'* — *'chairman'* may contribute to the

unified value of evaluated similarity. Now, consider another pair of terms: '*become*' — '*nominated*'. These terms probably share only a moderate degree of similarity in general, but a human reader will find their correspondence much more meaningful in this particular context. Identification of this context-dependent equivalence enables a reader to perceive that *John Daly* and *Hendrik Bokma*, respectively mentioned in the above texts, play an analogous part of being appointed to a managerial position. Existing similarity evaluation methods do not consider such analogies and do not provide tools for pointing them out.

Unlike common methods in automated natural language processing, cognitive research has emphasized the role of analogy in human thinking. The ability to detect analogous similarities between complex objects is presented by cognitive theories as a foremost mechanism in reasoning, problem solving and in human intelligence in general. The *structure mapping* theory (Gentner, 1983) presents analogy as mapping between two distinct systems. Particular entities that compose each system are not similar in general, but rather the relations among them resemble each other. Hence, entities in one system are perceived as playing a similar role to that played by corresponding entities in the other system. Another approach (Hofstadter et al., 1995) emphasizes the context-dependent interplay between perceiving features of the systems under comparison and creating representations that are suitable for mutual mapping.

Motivated by the above considerations, we present an initial step towards identifying automatically corresponding entities in natural language texts. At this stage of our research, correspondences are based on term similarity only, so terms describing similar topics are coupled. Identification and mapping of both entities and relations, using additional information, such as syntactic constructs (a direction which has been proposed in Hasse, 1995), will be handled in subsequent stages. However, presenting context-dependent topic correspondences in a pair of texts is by itself a non-trivial elaboration of standard approaches to document similarity.

Unsupervised specification of precise structure, let alone the optimal structure, is known to be an ill-posed problem even in classical tasks such as straightforward clustering. Nevertheless, we observe that our task here is to find relevant structure in the data. In section 2, we present a model for the structure we aim at. Then, in section 3, we recourse to a standard mechanism of capturing the quality of the proposed structure by a suitable cost function, followed by an algorithm seeking to minimize the cost. As in more studied learning-tasks, alternative costs or optimization methods are possible and form legitimate subject for future research. At this stage, we concentrate on demonstrating the feasibility of getting sub-topic similarity maps between text fragments through a novel bipartite clustering setting.

2. The Model: Term Subset Coupling by Bipartite Clustering

The present study suggests a framework for identifying corresponding sub-topics within a pair of text fragments. Our model represents sub-topics as groups of related terms, which are assumed to correspond to actual sub-topics. For this, the sets of terms appearing in each one of the fragments are divided into coupled subsets. A pair of coupled subsets, one from each fragment, is supposed to represent corresponding sub-topics from the compared fragments.

For the illustration, consider the following small term sets:

```
(1) {attendant, minister, government}
(2) {employee, manager}
(3) {student, university}
```

Term-subset coupling, based on semantic term similarity, applied to the first two term sets, might produce the following subset couples:

```
{attendant} — {employee}
{minister, government} — {manager}
```

For similar considerations applied to sets (1) and (3), the result might look like:

```
{attendant, minister} — {student}
{government} — {university}
```

These illustrative examples demonstrate expected topical partitions of the term sets according to the *diagnosticity principle* (Tversky, 1977). How each set is divided depends on how terms of both sets resemble each other: in the first case, the grouped topics are "*workers*" and "*management*"; in the second case – "*individuals*" and "*institutions*".

For obtaining subset coupling, we apply clustering methods. Quite a few previous works investigated the idea of identifying semantic substances with term clusters. Term clustering methods are typically based on the statistics of term co-occurrence within a word window, or within syntactic constructs (e.g. Pereira et al., 1993). The notion *pairwise clustering* refers to clustering established, as in the present study, on previous assessment of term similarity values – a process often based by itself on term co-occurrence data (e.g. Lin, 1999).

A standard pairwise clustering problem can be represented by a weighted graph, where each node stands for a data point and each edge is weighted according to the degree of similarity of the nodes it connects. A (hard) clustering procedure produces partition of the graph nodes to disjoint connected components forming a *cluster configuration.*

Our setting is special in that it considers only similarity values referring to term pairs from two distinct text fragments, such as '*attendant*'–'*manager*' in the example above, but not '*attendant*'–'*minister*'. The exclusion of within-fragment similarities is conformed to our context-oriented approach, but there is no essential restriction on incorporating them in a more comprehensive model. Consequently, our setting is represented by a *bipartite graph* containing only edges connecting nodes from two distinct node sets, each of which associated with terms from a different text fragment. A term that appears in both articles is represented independently in both sets.

The use of clustering within a bipartite graph (*bipartite clustering*) is not common in natural language processing. Hofmann and Puzicha (1998) introduce taxonomy of likelihood-based clustering algorithms for co- occurrence data, some of which produce bipartite clustering. To illustrate their soft clustering method, they present sets of nouns and adjectives that tend to co-occur in a large corpus. Sheffer–Hazan (1997) developed a bipartite clustering algorithm based on description length considerations for purposes of knowledge summarization and text mining. Both works exploit co-occurrence data for exposure of global characteristics of a corpus. The present study refers too, through its use of pre-compiled similarity data, to co-occurrence statistics in a corpus. Here, we go beyond that to get fine-grained context-dependent groupings in the term sets of particular text-fragment pairs.

When pairwise clustering algorithms are applied on a bipartite graph, the assignment of a term from one of the sets into a cluster is influenced by the assignments of similar terms from the other set. Each one of the resulting clusters, if contains more than a single element, necessarily contains terms from both parts, e.g. <*minister, government, manager*> in the example above. Therefore, a cluster couples two term subsets, each from a different fragment: the subset {*manager*} is coupled to the subset {*minister, government*}. Clusters containing a single element represent terms that could not be assigned to any of the coupled subsets by the clustering method.

3. Algorithms: Balancing Within-Cluster and Between-Cluster Similarities

Let X and Y denote the sets of the terms appearing in a given pair of articles. We currently use the "*bag of words*" model, where term repetitions are not counted. Non-negative similarity values, $s(x,y)$, are given (as input) for each $x \in X$ and $y \in Y$. Assume that some clustering procedure is applied to the appropriate bipartite graph, so that a partition of the graph nodes is given. Denote by Cx the part containing $x \in X$. Recall that if Cx contains additional elements, some of them must be elements of Y. Hence, Cx represents coupling of the subsets $X \cap Cx$ and $Y \cap Cx$.

A basic clustering strategy is the greedy *single-link* agglomerative method. It starts with a configuration in which for each $x \in X$ and $y \in Y$, $Cx = \{x\}$, $Cy = \{y\}$. Then, the method repeatedly merges a pair of clusters Cx and Cy such that x and y are the most similar elements for which $Cx \neq Cy$. The result is a hierarchical arrangement of clusters, also called *dendogram.* There is no fixed recipe of how to select the best clustering configuration (partitioning) in the hierarchy. Furthermore, in our case the number of target sub-topics is not known in advance. We thus refer to the obtained hierarchy as representing a range of possible cluster configurations, corresponding to varying granularity levels.

An alternative approach states in advance what is expected from a good clustering configuration, rather than letting the merging process dictate the clustering as in the case of single-link. This is customarily done by formulating a *cost function*, to be minimized by an optimal configuration. In our case, as in clustering in general, a cost function reflects the interplay between two dual constraints:

(i) *Maximizing within-cluster similarity*, i.e. the tendency to include similar objects in the same cluster. It should be stressed that in the bipartite setting the notion of 'within-cluster' refers to similarity values between pairs of terms from coupled subsets, while the actual similarities within each subset are not considered. The excessive satisfaction of this constraint dictates a cluster configuration containing many small clusters, each characterized by high similarity values among its members.

(ii) *Minimizing between-cluster similarity*, i.e. the tendency to avoid assigning similar objects (in the bipartite setting – from distinct fragments) into different clusters. The excessive satisfaction of this constraint results in obtaining large clusters, so that only minimal between-cluster similarity is present.

We have considered several cost function schemes, reflecting different types of interactions between the above two constraints. One particular scheme, which enables obtaining context-dependent subset coupling at various granularity levels, is presented here.

This scheme captures the between-cluster similarity minimization constraint by including, for each term $x \in X$ (and correspondingly for each $y \in Y$), a cost component proportional to the between-cluster similarity values associated with that term, i.e. proportional to $\sum_{y \in Y-Cx} S(x,y)$.

According to the other constraint of within-cluster similarity maximization, each term x is supposed to be assigned into a cluster such that its contribution to the total measure of within-cluster similarity is maximal. To obtain a cost measure, which is inversely proportional to the contribution of x to total within-cluster similarity, we measure the total degree of within-cluster similarity obtained if x were removed from its cluster Cx. That is, we add for each $x \in X$ (and correspondingly for each $y \in Y$) a cost component proportional to the total contribution to within-cluster similarity of the other subset members: $\sum_{x' \in X \cap Cx-\{x\}} \sum_{y \in Y \cap Cx} S(x',y)$. This component is further multiplied by $1/|X|$ for normalizing it relatively to the entire set size.

Finally, the cost function scheme introduces a parameter, $0 < \alpha < 1$, which controls the relative impact of each of the two constraints. The resulting scheme is thus weighted sum of the two cost components for all terms in X and Y:

$E(M) =$

$$\sum_{x \in X} \left((1-\alpha) \sum_{y \in Y-Cx} s(x,y) + \alpha \frac{1}{|X|} \sum_{x' \in X \cap Cx-\{x\}} \left(\sum_{y \in Y \cap Cx} s(x',y) \right) \right) +$$

$$\sum_{y \in Y} \left((1-\alpha) \sum_{x \in X-Cy} s(y,x) + \alpha \frac{1}{|Y|} \sum_{y' \in Y \cap Cy-\{y\}} \left(\sum_{x \in X \cap Cy} s(y',x) \right) \right)$$

Varying α has the effect of changing cluster size within the optimal configuration, due to the varying impact of the two constraints (increasing α reduces cluster size, and vice versa). Another interesting property of this scheme is that coupling two singletons, which have a positive similarity value, always reduces the total cost. This is because such coupling, forming a two-member cluster, reduces between-cluster similarity cost and does not increase within-cluster similarity cost.

Note that *E(M)* pretends to reflect balance of constrains, as described above, only for a particular pair of documents at a time. Its potential value as a basis for unified document similarity measure, sensitive to context-dependent and analogous similarities, is yet to be investigated.

There are sophisticated techniques to compute an optimal solution minimizing the cost function for a given α value, e.g. simulated annealing and deterministic annealing (Hofmann and Buhmann, 1997). A simple strategy, assumed to suffice for preliminary demonstration of cost function behavior for any α, is a greedy method, similar to the single-link method. It starts with a configuration in which for each x and y, $Cx = \{x\}$, $Cy = \{y\}$ and then merges repeatedly the two clusters whose merge minimizes the cost mostly. Unlike single-link, this process stops when no further cost reduction is possible.

4. Results: Hierarchy and Granularity

Our experiments were performed for term coupling between pairs of Reuters news articles.

Here we qualitatively demonstrate the results using the same pair of articles of the example in Section 1 (devising a quantitative evaluation method for our task is an issue for future research). We used pairwise term similarity values that were compiled by Dekang Lin, using a similarity measure based on information theoretical considerations, from co-occurrence data in a large corpus of news articles (Lin, 1999; data available for download from http://www.cs.umanitoba.ca/~lindek/sims.tgz).

The term sets were taken to be the sets of words, in each article, which had at least one positive similarity value with a term in the other article. The vocabulary included verbs, nouns, adjectives and adverbs, excluding a small set of stop words (e.g. '*about*'). The *Conexor NP&Name* Parser (Voutilainen, 1997) was used to obtain word lemmas.

Figure 1 displays detailed term subset coupling generated by the single-link procedure. The hierarchy is indicated by different widths of contours bounding term subsets. Each contour width presents the clusters obtained after all merges that were imposed by similarity values larger than a threshold t. Coupling connections are displayed for the most detailed granularity level. An apparent drawback of this method is that many terms are assigned into clusters only in a late stage, although they seem to be related to one or more of the smaller clusters. E.g. '*management*' seems to be related, and indeed has non-zero similarity values, to '*chairman*' '*director*' and '*president*' as well as to '*chief*'. This is indicated by including such terms in the largest thin frames in Figure 1, but not in any bold smaller frame.

We have also implemented more sophisticated methods proposed recently (Blatt et al., 1997; Gdalyahu et al., 1999) that are related to the single-link strategy. These methods are designed to overcome cases where few "noisy" data points invoke union of clusters that would have remain separate in the absence of these points. Both methods repeatedly sample stochastic approximated cluster configurations. Elements, persistently found in the same cluster across the sample, are assigned to the same cluster also in the final solution. The results obtained with these methods are qualitatively similar to those obtained with single-link. This suggests that the fact that certain terms remain uncoupled in high granularity levels can not be attributed to random inaccuracies in the data.

Figure 2 displays a detailed term subset coupling generated by the cost-guided greedy strategy. The lack of strict hierarchy prevents displaying a wide range of granularity levels within the figure, so a sample of clusters is presented. The gray clusters demonstrate the impact of lower α values on cluster granularity. Several of the coupled term-subsets represent actual sub-topics, such as "*trade operations*" and "*managerial positions*". Comparing with the single link algorithm, the cost-based algorithm does succeed to couple related terms such as '*management*' and '*chairman*' within a relatively tight cluster. Note also that the algorithm couples the words '*become*' and '*nominate*', as discussed in Section 1.

5. Conclusions

This paper describes a preliminary step, suggesting bipartite term coupling as an attractive approach for detecting sub-topic correspondence. Future work is required to investigate aspects that have already been mentioned, such as the use of other similarity measures, the incorporation of within-document similarities and additional search strategies. Another direction we are considering is integration of data from several sub-topic maps, in order to modify the original term similarity matrix, starting an iterative algorithm in the EM style. In addition, we wish to study how additional attitudes to clustering, e.g. the one described by Pereira et al. (1993), are related to our setting. It is also necessary to develop a quantitative evaluation method, possibly based on comparing the performance of our method with that of human subjects in similar tasks.

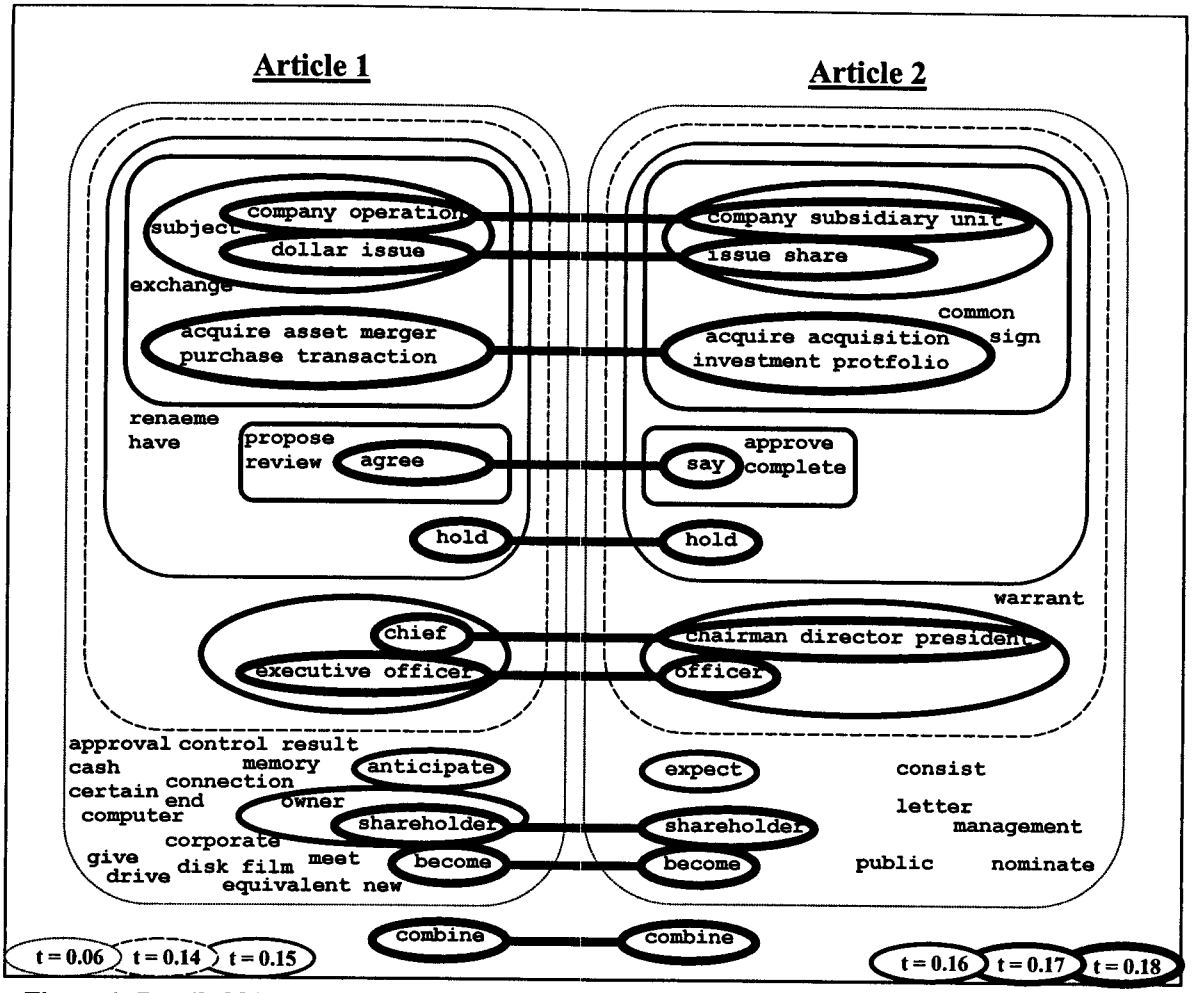

Figure 1: Detailed hierarchical subset coupling in a pair of Reuters news articles, as generated with the single link method. Each of the presented merging stages is characterized by the similarity value t that imposed the merge at that stage. The different stages are indicated by different contour widths. Coupling connections are indicated as straight lines and are displayed only for the most detailed level t = 0.18.

From a broader perspective, this research initiates an original unsupervised learning framework, capturing similarity of complex objects. It is hoped that future results will provide a significant contribution to both setting and achieving information technology tasks, so they better reflect human thinking and needs.

Acknowledgements

We thank Dekang Lin for the use of his Automatically Generated Thesaurus. This research was supported by ISRAEL SCIENCE FOUNDATION founded by The Academy of Sciences and Humanities (grant 574/98-1).

References

Blatt M., Weisman S., and Domany E. (1997) Data Clustring Using Model Granular Magnet. *Neural Computation*, 9/8, pp. 1805–1842.

Deerwester S.,. Dumais S. D., Furnas G. W., Landauer T.K., and Harshman R.A. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41/6, pp. 391–407.

Gdalyahu Y., Weinshall D., and Werman M. (1999) A Randomized Algorithm for Pairwise Clustering. In *"Advances in Neural Information Processing Systems 11 (NIPS*98)"*, M. S. Kearns, S. A. Solla & D. A. Cohn, ed., MIT Press, Boston (to appear).

Gentner, D. (1983) Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7/2, pp. 155–170.

Haase K. (1995) Analogy in the Large. In *"IJCAI-95: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence" Vol. 2*, Montreal, pp. 1375–1380.

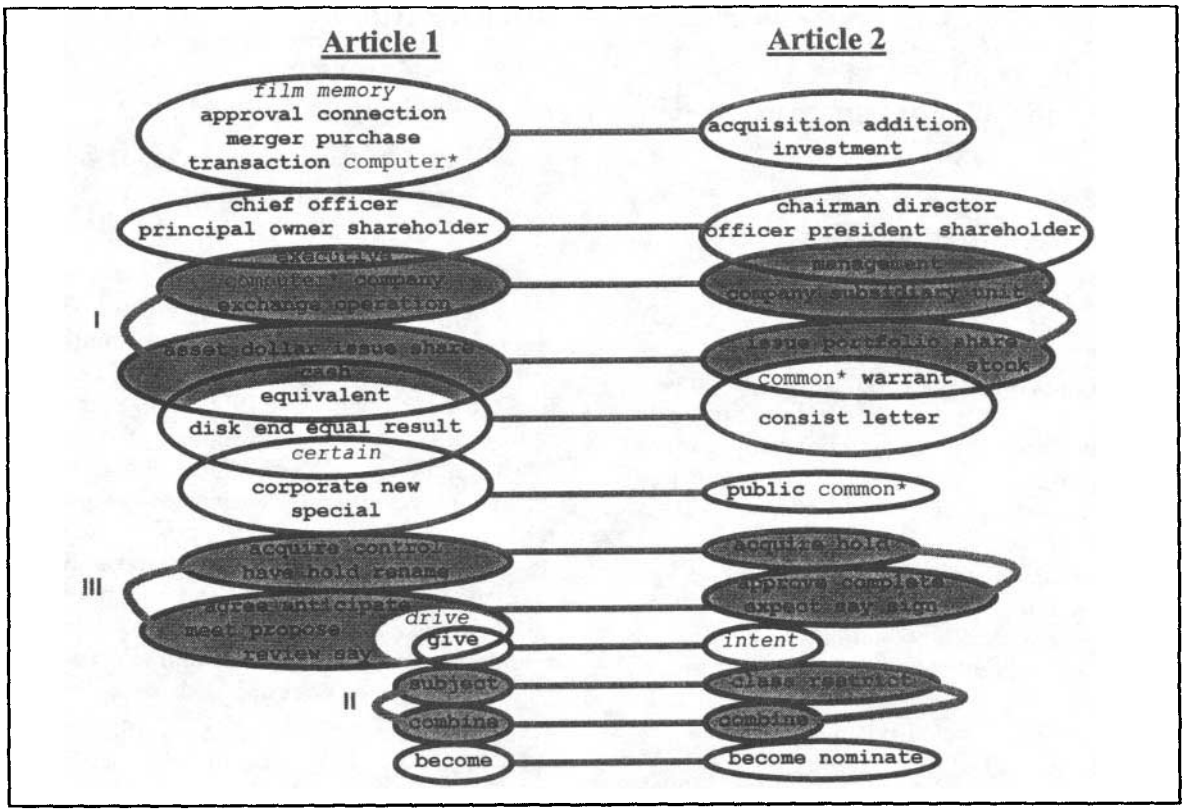

Figure 2: A sample of subset coupling output as generated with a greedy method applied to the cost-based scheme for α values between 0.53 to 0.73. Coupling connections are indicated as straight lines. The gray areas indicate coupled subsets of lower granularity level, which were formed for $0.53 \leq \alpha \leq 0.68$ (I), $0.53 \leq \alpha \leq 0.55$ (II) and $\alpha = 0.53$ (III). Terms that swapped their membership for different α values appear, whenever possible, in the intersection of the appropriate subsets; otherwise, they are marked with asterisk. Terms that for some α value were not included in any coupled subset are in Italics.

Hofmann T. and Puzicha J. (1998) Statistical Models for Co-occurrence Data. *AI Memo No. 1625 / CBCL Memo No. 159*, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Massachusetts Institute of Technology, Boston, 21 p.

Hofmann T. and Buhmann J. (1997) Pairwise Data Clustering by Deterministic Annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19/1, pp 1–14.

Hofstadter D. R. and the Fluid Analogies Research Group (1995) *Fluid Concepts and Creative Analogies*. Basic Books, New-York, 518 p.

Lin D. (1999) Automatic Retrieval and Clustering of Similar Words. In "*Proceedings of the Seventeenth International Conference on Computational Linguistics and the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics (COLING–ACL '98)*", Montreal (to appear).

Kowalski G. (1997) *Information Retrieval Systems: Theory and Implementation.* Kluwer Academic Publishers, Boston, pp 125–148.

Pereira, F. C. N., Tishby N. Z., and Lee L. J. (1993) Distributional Clustering of English Words. In "*Proceedings of the Thirty-First Annual Meeting of the Association for Computational Linguistics (ACL)*", Columbus, OH, pp 183–190.

Sheffer–Hazan S. (1997) *Knowledge Discovery and Summerization by Bipartite Graph.* Unpublished Master`s Thesis (in Hebrew). Bar-Ilan University, Ramat-Gan, Israel, 68 p.

Tversky A. (1977) Features of Similarity. *Psychological Review*, 84/4, pp 327–352..

Voutilainen A. (1997) The Conexor NP&Name Parser (ENG-CG). Web page: http://conexor.co.helsinki.fi/NPintro.html.