Towards Visual Exploration of Topic Shifts

Kilian Thiel, Fabian Dill, Tobias Kötter, and Michael R. Berthold

Abstract—This paper presents two approaches to visually analyze the topic shift of a pool of documents over a given period of time. The first of the proposed methods is based on a multi-dimensional scaling algorithm, which places vectors representing terms occurring in certain years (period-frequency-vectors) in a spatial, two-dimensional space. This kind of visualization enables the detection of terms occurring in documents, published in particular years, or terms spread over different years. The second method uses a graph based approach. Publishing dates of documents, as well as their terms are represented by the vertices of a graph. Terms related to a specific publishing year are connected to the vertex of the year via an edge. By usage of activation spreading techniques, terms frequently occurring in documents published in particular years can be discovered visually. We tested both approaches with 2431 abstracts of papers published in the IEEE Transactions on SMC-A, SMC-B, and SMC-C in the years 1996 to 2006. Our experiments indicate that a number of interesting terms can be nicely separated in clumps according to individual years or periods of time. In addition, one can visualize the emergence of specific terms over certain periods of time and how these and other terms fade away again later.

I. INTRODUCTION

Publishers of scientific publications, computer linguists and many others find it interesting and often extremely valuable to know if and how the topics of their publications are changing and shifting over time. It is interesting, e.g. to detect which topics become untended over time, completely disappear and which are coming up newly. Also of prime interest is the detection of trends or transformations of a particular vocabulary. In order to achieve this, vast amounts of documents have to be analyzed in such a way that topics, relevant terms, or concepts can be revealed, as well as temporal changes of these. To complete this task manually simply would not be feasible in a reasonable amount of time if the number of documents is large. This means that automatic methods have to be found which are able to handle this task. In this paper, two methods are introduced aiming to visualize this kind of topic shift.

The first of these methods is based on a visualization of relevant terms contained in the documents by means of a multi-dimensional scaling algorithm (MDS). In [1], concepts of documents are already visualized by a visualization of similarities (VOS) method, which is similar to a MDS visualization [2]; both methods try to preserve the proximity information of objects, when mapping them onto a usually two-dimensional space. The main difference to [1] is the

Kilian Thiel, Fabian Dill, Tobias Kötter, and Michael R. Berthold are with the ALTANA-Chair for Bioinformatics and Information Mining, University of Konstanz, 78457 Konstanz, Germany `thiel@inf.uni-konstanz.de`, `berthold@inf.uni-konstanz.de`

representation of concepts or terms and their extraction. In our approach, the terms are extracted directly from the documents without the use of a thesaurus. The extracted terms are represented by period-frequency-vectors, explained in Section IV, and the distances between these vectors are then visualized.

The second approach is based on a network model or graph, where the vertices represent information units such as authors, relevant terms regarding to a document, or the documents itself [3]. These vertices are connected via weighted edges to each other according to their relatedness. This can lead to big networks if many documents are analyzed. Since, in this paper, we focus on the shift of topics or concepts and the change of vocabulary, only this information is taken into account when creating such a network. This means that the units of information represented by the vertices are the most relevant terms as well as the publishing dates of the documents. Edges exist only between vertices representing terms and those representing publishing dates. To explore these networks, we use activation spreading techniques ([5], [6]).

The paper is organized as follows: in the next Section, we briefly introduce the data which we used to test both approaches, as well as the preprocessing of the documents. In Section III, the multi-dimensional scaling method is described. In Section IV, the creation and the visualization of vectors representing terms occurring in certain years (period-frequency-vectors) is explained. We subsequently explain in Section V the usage of the network model by means of exploration of the vocabulary occurring in particular periods and its shift.

II. DATA

To test our approaches for detecting topic shifts and change of vocabulary in documents, we use abstracts of scientific papers, since these abstracts can be accessed using bibliographic information services. In particular, we used abstracts and citation information of papers published in the IEEE Transactions on SMC-A, SMC-B, and SMC-C in the years 1996 to 2006. The data consist of 2431 abstracts, which translates to roughly 220 abstracts per year.

A. Preprocessing and term extraction

After parsing the abstracts, the text has to be tokenized. Additionally, part-of-speech (POS) tags are assigned to the tokens to be able to filter out particularly uninteresting parts-of-speech, such as articles, and keep more relevant parts like nouns, verbs and adjectives. The software used to tag English

words is an implementation of a log-linear part-of-speech tagger using the Penn Treebank POS tag set ([7], [8]).

1) Filtering: Three different filters are applied to exclude irrelevant characters and words: a punctuation filter which filters out punctuation marks, a stop word filter to filter out non informative words like "a", "an", "and", "or" etc., as well as a POS tag filter to filter out irrelevant parts-of-speech like determiner.

2) Stemming: In order to avoid treating words differently which occur in different flexions, e.g. verbs with different forms of conjugation or nouns with different forms of declension, all different flexions have to be detected and unified. To achieve this, stemming, which reduces the inflected words to their stem or base [9], is the appropriate method. Since we are dealing with English words only, we used the Porter stemmer algorithm [10].

3) Term extraction: The final step after filtering and stemming is the extraction of relevant terms of each document. Classifying a term as relevant can be done in several ways ([12], [13]). Here, the relative term frequency (tf), in combination with the inverse document frequency (idf), is used to identify the relevance of each term. The relative term frequency of a term t in document d is specified as follows:

$$\text{tf}(t,d) = \frac{f_d(t)}{a(d)},$$

with $f_d(t)$ as the frequency of term t in document d and $a(d)$ as the number of terms in d. The inverse document frequency of t is computed by:

$$\text{idf}(t) = \log\left(1 + \frac{N_D}{f_D(t)}\right),$$

where N_D specifies the number of all documents D and $f_D(t)$ the number of documents containing term t.

To specify and extract the relevant terms with the help of these frequencies the tfidf value can be computed by simple multiplication of the two terms:

$$\text{tfidf} = \text{tf} \cdot \text{idf}.$$

Terms with a low tfidf value will occur in most of the documents (small idf value) and/or they occur only rarely in the documents they are contained in (small tf value). By filtering them out, terms with average frequency are maintained. According to Zipf [11], these terms are most relevant.

Subsequently, a threshold can be chosen, which determines if a term is classified as relevant or not, concerning to a document. On the other hand, a maximum number k of terms to extract can be specified and the k terms with the largest tfidf values are extracted.

The 2431 abstracts including titles were tokenized into 21279 tokens, after tagging, stemming and filtering 18160 different words (stems) remained. We decided to extract 5000 terms with the largest tfidf value out of these words, which is roughly twice the number of documents. This means that in average approximately two terms per document are extracted.

In Section IV-A the reason why we haven chosen 5000 terms is described more detailed.

The extraction of terms is an essential point in the detection of topic shifts, since here a topic is represented by extracted terms. If too few terms are extracted, many topics will be ignored. Further it is possible that whole documents will be ignored since no terms are extracted from them due to their low tfidf value. On the other hand, if only a small number of topics is of interest, a low k value can be a way to achieve this goal, at the cost of ignoring many terms and topics. If the k value is too large, too many irrelevant terms are extracted, complicating the detection of essential topics and their shift. The topic detection used here is a first attempt and leaves room for improvement, e.g. by applying extracted concepts as topics instead of terms.

III. MULTI-DIMENSIONAL SCALING

The projection of objects of high-dimensional spaces onto two or three dimensions causes a loss of proximity information. Multi-dimensional scaling (MDS) methods [14] try to preserve the pairwise distance between objects when mapping them to lower dimensional space by minimizing an appropriate error function. This means that the proximity information between objects is kept as accurately as possible. The reduction of dimensions allows the visualization of high-dimensional points in a lower-dimensional space.

The Sammon algorithm [15], is one of the best known multi-dimensional scaling methods, which computes such a mapping. For each object of the high-dimensional space $\vec{X}_i \in \mathbb{R}^H$, a spatial representation $\vec{x}_i \in \mathbb{R}^L$ in the lower-dimensional space (usually $L = 2$) has to be found ($1 \le i \le N$, with N as the number of objects).

To maintain the proximity information between objects as close as possible, the individual positions and thereby the distances of two objects $\vec{x}_i$, $\vec{x}_j$ in the low-dimensional space $d_{ij} = d(\vec{x}_i, \vec{x}_j)$ have to be approximated to the distances of two objects $\vec{X}_i$, $\vec{X}_j$ in the high-dimensional space $D_{ij} = D(\vec{X}_i, \vec{X}_j)$, which is:

$$\forall_{i \neq j} : D_{ij} \approx d_{ij}, 1 \le i, j \le N.$$

Usually the Euclidean metric is used to measure these distances:

$$D_{ij}^2 = \sum_{q=1}^{H}(X_{i,q} - X_{j,q})^2,$$

$$d_{ij}^2 = \sum_{k=1}^{L}(x_{i,k} - x_{j,k})^2.$$

To approximate the distances as best as possible, Sammon formulates a minimization problem of a cost function, which aggregates the weighted squared differences of the distances:

$$E = \sum_{i=1}^{N}\sum_{j>i}^{N} w_{ij}(d_{ij} - D_{ij})^2.$$

For each point $\vec{x}_i$ that is randomly initialized, a steepest gradient method is applied at each step to iteratively minimize
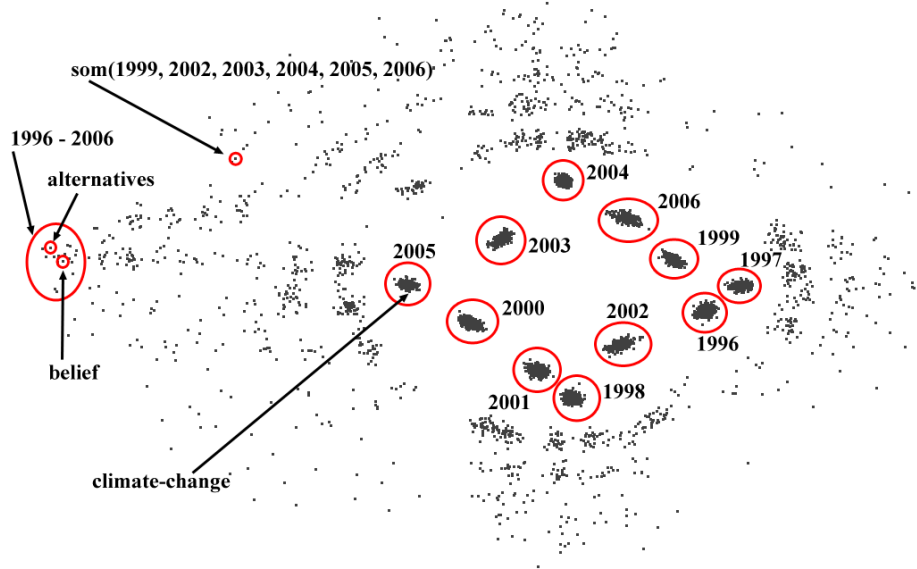
Fig. 1. Each of the 5000 dots represents a term described by a 11-dimensional period-frequency-bit-vector. Terms which appear in clumps, here surrounded by circles, occur in particular periods. A few noteworthy terms are indicated by arrows.

the remaining cost E. Usually several iterations are needed by the algorithm, to converge to a local cost minimum.

IV. PERIOD-FREQUENCY-VECTORS

To apply multi-dimensional scaling to visualize the distribution of terms over the time periods and gain information about the shift of topics, the extracted terms are represented by period-frequency-vectors. A period-frequency-vector of a term is a vector consisting of frequencies of that term according to particular periods like years or months. For each period $p_i \in P$ of the set of all periods $P = \{p_1, ..., p_m\}$, the frequency $f_{p_i}(t)$ $(1 \leq i \leq m)$ of term t is determined by the aggregation of all occurrences of t in those documents, which are published in period p_i. In the resulting period-frequency-vector $\vec{fp}_t$, position i is set to this frequency value, i.e.:

$$\vec{fp}_t = (f_{p_1}(t), ..., f_{p_m}(t)).$$

In this setting the dimension of the period-frequency-vectors is 11 derived from the 11 years from 1996 to 2006.

A. Visualization of the period-frequency-vectors

The extracted terms represented by the period-frequency-vectors defined above can be visualized by the multi-dimensional scaling method described in Section III.

The aim is to map the 11-dimensional period-frequency-vectors via MDS to two-dimensional objects, which can be displayed in a scatterplot. Terms occurring in a particular period only should appear close together in a clump well separated from terms occurring in other periods. For the visualization by means of MDS the important information of a term's period-frequency-vector is the occurrence of a term in a particular period only, not its frequency. A different frequency of terms occurring in the same period of time only would result in a distance value greater than zero. This means

that the visualized terms would not appear close together. Hence, the vectors are modified to bit-vectors, indicating whether a term occurs in a period or not. A vector $\vec{fp}_t$ representing term t

$$\vec{fp}_t = (f_{p_1}(t), ..., f_{p_m}(t))$$

is modified to the following bit-vector:

$$\vec{fp}_t^{\,*} = (g_{p_1}(t), ..., g_{p_m}(t)),$$

with

$$g_{p_i}(t) = \begin{cases} 1 & \text{if } f_{p_i}(t) > 0 \\ 0 & \text{else.} \end{cases}$$

In Figure 1 all 5000 terms are visualized, described by period-frequency-bit-vectors. Each dot displayed in the scatterplot represents one term. Terms appearing in the same period of time only are displayed close to each other, forming a clump. The 11 obvious clumps of Figure 1 are emphasized by the circles surrounding them. They represent each of the years from 1996 to 2006. The majority (3547) of the 5000 different terms can be assigned to one of these 11 clumps alone. This means that the majority of the topics represented by the extracted terms changes from year to year and less than the half of the topics have been discussed over more than one year.

The 11 clumps are positioned approximately in the middle of Figure 1. This is due to the fact, that the period-frequency-vectors of terms of different clumps all have the same distance ($\sqrt{2}$) to each other, since they occur in only one year. Therefore, the final position of a clump inside the middle of the figure depends on the initialization of the 2-dimensional points which is done by random. The year in which the terms of a clump are published has no effect on that. The period-frequency-vectors of terms occurring in more than one year are placed in ovals around the clumps

since their distance to terms in the middle is getting larger the more years they appear in. Terms occurring in all 11 years are placed together at the left outer side of the figure. These terms build the 12th "clump" consisting of only 14 terms.

As expected, these terms are basically very general as they occur in many documents published in all of the years, in contrast to terms of the other 11 clumps which are more specific. For instance words like "alternatives", "belief", "algorithm" or "knowledge" are part of the 12th clump.

On the other hand, there are words which occur only in specific years, e.g. "human-robotic", "biometric", "wavelet-based" or "climate-change". The latter is contained in a document about a model to predict climate-change impact on fish catch [16], published in the year 2005. In the former years, this term did not occur in any of the analyzed documents, which leads to the assumption that the interest of that topic in these periods was not that strong. Nowadays, in the year 2007, the climate-change debate is still under way. It remains to be seen, if the newly arisen debate on climate change will be reflected in the SMC publications 2007 or later.

However, Figure 1 clearly shows that there is a change of vocabulary and a shift of topics over the analyzed years. Additionally, it can be seen that only $\approx 0.28\%$ (14 terms) of the extracted vocabulary is used during all the years considered here.

Further experiments showed that a reduction of the number k of extracted terms to 2500 or less leads to the complete disintegration of the 12th clump while the other 11 clumps remain. The idf value of most of the terms occurring in documents spread over all periods of time is lower than the value of most of the terms occurring only in documents of a certain year. This means that a small idf value leads to a lower tfidf value. When the k value is decreased these terms will be filtered out first. On the other hand, the increase of the k value implicates an increase of the terms assigned to the 12th clump, which are mostly very general. Since we wanted to show that there is a 12th clump but also keep it nearly as small as possible to avoid extraction of the general terms, we have chosen a k value of 5000.

V. NETWORK APPROACH

Another way to model the relations of documents is via a network or graph, where the information entities, like authors, terms, titles, or documents as a whole are represented as vertices and the relations between them, e.g. citations, authorship, containedness, or term cooccurrences as edges [3]. Additionally, edges can be weighted, indicating the relevance or strength of the modeled relation.

The query processing and the exploration of these networks is usually done by spreading activation methods such as Branch-and-Bound search or Hopfield activation [4]. In spreading activation methods the vertices representing the terms of a query obtain an initial activation energy. This step is also called priming. The energy spreads across the network by following the outgoing edges, thereby activating

the visited vertices with a part of the energy, depending on the weight of the edges. The activation spreading is stopped if either a certain number of activated vertices is reached, or the activation energy falls below a specified threshold.

A. Visualization of topic shifts using network graphs

In our second approach we utilized the network described above to explore and visualize the change of vocabulary and the shift of topics. The extracted terms as well as the 11 years from 1996 to 2006 are modeled as information units, i.e. vertices of the graph. The only kind of relationship represented as edges in the graph is the occurrence of a term in a certain year, i.e. a term occurring in a document published in a certain year establishes an edge between the year and the term. To obtain all the terms occurring in a particular year, the vertex of this year has to be primed, such that the activation is spread to adjacent vertices which represent the wanted terms.

Figure 2 and Figure 3 visualize such a graph consisting of 511 vertices representing 500 terms and 11 years. Due to reasons of space we chose to display only a small graph. However, the implemented graph layout enables to efficiently visualize about 1000 to 2000 of vertices on a high-resolution display. In order to reduce the number of vertices we extracted only 500 terms with the largest tfidf values. Additionally, terms which occur less than three times in a document are filtered out. As already described in section II-A.3 the restrictive filtering of terms has disadvantages like only considering terms with the highest average frequency and ignoring other terms.

The layout of the graph places the vertices representing years in a shell around the center. Those terms which are connected to one year only are placed in an arc around the vertex of the corresponding year on the outside of the shell. Terms which occur in more than one year are placed in the inside of the shell. A force-directed layout pushes the terms towards the vertices of the corresponding years, leading to a positioning in the geometric center of them.

Figure 2 shows the graph with 500 terms and 11 years from 1996 to 2006. Positively activated vertices are visualized by black circles, negatively activated by white circles and vertices without activation are colored gray. In this graph the years 2004, 2005, and 2006 are positively primed and 1996 and 1997 are negatively primed. As one can see terms like "regression", "fuzzy-logic", "fuzziness" and "fuzz"[1] etc. only occur in 1996 and 1997. More recently, terms like "svm", "clustering", "self-organizing", "gene" and "e-business" appear in 2004, 2005 and 2006. The reason that the terms "fish" and "swarm" emerge in 2004 and 2005 could be related to "climate-change" occurring in [16], which was found to be frequent in 2005 with the MDS approach described above. The gray vertices between the referring years represent more general terms frequently used in both sets of years, for example "synchronization", "hierarchy", and "discrimination".

---

[1]Note that the terms represent the word stems only, hence the truncation.

In Figure 3 the same graph as in Figure 2 is displayed. The older years, ranging from 1996 to 2001, were primed negatively and the recent years from 2002 to 2006 were primed positively. Those terms occurring in both sets of years are activated positively and negatively which neutralizes the activation, hence they are colored gray. This setting clearly separates the inner shell into three regions: old, general, and recent terms, colored in white, gray and black, respectively.

As described above, "old" terms like "fuzzy-logic" etc. can be detected in the years 1996 to 2001. Further it can be seen that more recent topics from the field of bioinformatics like "medical", "tumor" "gene" and "fingerprint" appear in 2003 and 2005.

VI. CONCLUSIONS

In this paper two methodologies were presented to visualize topic shifts and change of vocabulary of scientific papers. The first approach represented the preprocessed terms as period-frequency-vectors. These vectors are mapped onto a two-dimensional space, using a multi-dimensional scaling method and displayed in a scatterplot. This kind of visualization enables the detection of clumps of terms, which occur only in particular periods or over various periods.

The second approach is based on a network model. The extracted terms as well as the periods are represented by vertices of a graph. The terms occurring in a certain period are connected to the referring vertex. Via spreading activation methods terms occurring in a certain period can be found by priming the period vertex. Using an appropriate layouting mechanism, this approach nicely visualizes the topic shift and a change of vocabulary.

Whereas the first method allows the visualization of many ($>$10,000) terms, the second method is more restricted by the available space. However, the terms themselves and the overlap between several periods are perceived more conveniently and clearly with the network approach, since it supports interactive user exploration. Therefore this method can be favored if an appropriate method of topic extraction exists which extracts only a smaller number ($<$1,000) of topics out of a document set.

In both methods the extraction of topics from a document set is done by simple term extraction. After preprocessing, the k terms with the largest tfidf values are filtered out and taken as topics. This way of topic extraction leaves room for improvement, since the right choice of the number k of terms has a great impact on the extracted topics. Also the assignment of terms as topics can be improved by using i.e. extracted concepts as topics.

Quite obviously much work remains to be done. Currently only a limited amount of information is included in both methods, for instance the sharp cutoff threshold for relations to be included poses a problem. It will also be interesting to see how more interactive techniques will allow exploration of regions of interest, especially in the network based visualization. It will also be of interest to include more information in addition to simply the year-term links and allow also the exploration of connections between words and more general concepts, such as derived from an ontology or a thesaurus.

REFERENCES

[1] N.J. van Eck, L. Waltman, J. van den Berg, and U. Kaymak. Visualizing the computational intelligence field. *IEEE Computational Intelligence Magazine*, 1(4):6-10, 2006

[2] N.J. van Eck, and L. Waltman. VOS: A new method for visualizing similarities between objects, *Proc. 30th Ann. Conf. German Classification Society*, 2006

[3] R. K. Belew. Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. In SIGIR '89: *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11-20, New York, NY, USA, 1989. ACM Press

[4] K.L. Kwok. A neural network for probabilistic information retrieval. In SIGIR '89: *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2130, New York, NY, USA, 1989. ACM Press

[5] H. Chen, K. Basu, and T. Ng. An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound search vs. connectionist hopfield net activation. *Journal of the American Society for Information Science*, 46(5):348369, 1995

[6] G. Salton, and C. Buckley. On the Use of Spreading Activation Methods in Automatic Information Retrieval. In SIGIR '88: *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 147–160, New York, NY, USA, 1988, ACM Press

[7] K. Toutanova, and C.D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, 2000.

[8] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *In Proceedings of HLT-NAACL*, pages 252-259, 2003

[9] R. Ferber. Information Retrieval Suchmodelle und DataMining Verfahren für Textdammlungen und das Web. dpunkt.verlag, first edition, 2003

[10] M.F. Porter. An algorithm for suffix stripping, *Program*, 14(3) pp 130-137, 1980

[11] G.K. Zipf. Human behaviour and the principles of least effort, Addison-Wesley, 1949

[12] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. Text mining at the term level. In *Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, page 65, 1998.

[13] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word cooccurrence statistical information. *Int. J. AI Tools*, Vol. 13, 157169, 2004.

[14] T.F. Cox, and M.A. Cox. Multidimensional Scaling. *Monographs on Statistics and Applied Probability*, Chapman and Hall, 1994

[15] J.W. Sammon, Jr. A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, C18(5):401-409, May 1969

[16] B. K. Biswas, Y. M. Svirezhev, B. K. Bala. A model to predict climate-change impact on fish catch in the world oceans. *IEEE Transactions on Systems, Man, and Cybernetics*, Part A 35(6): 773-783, 2005
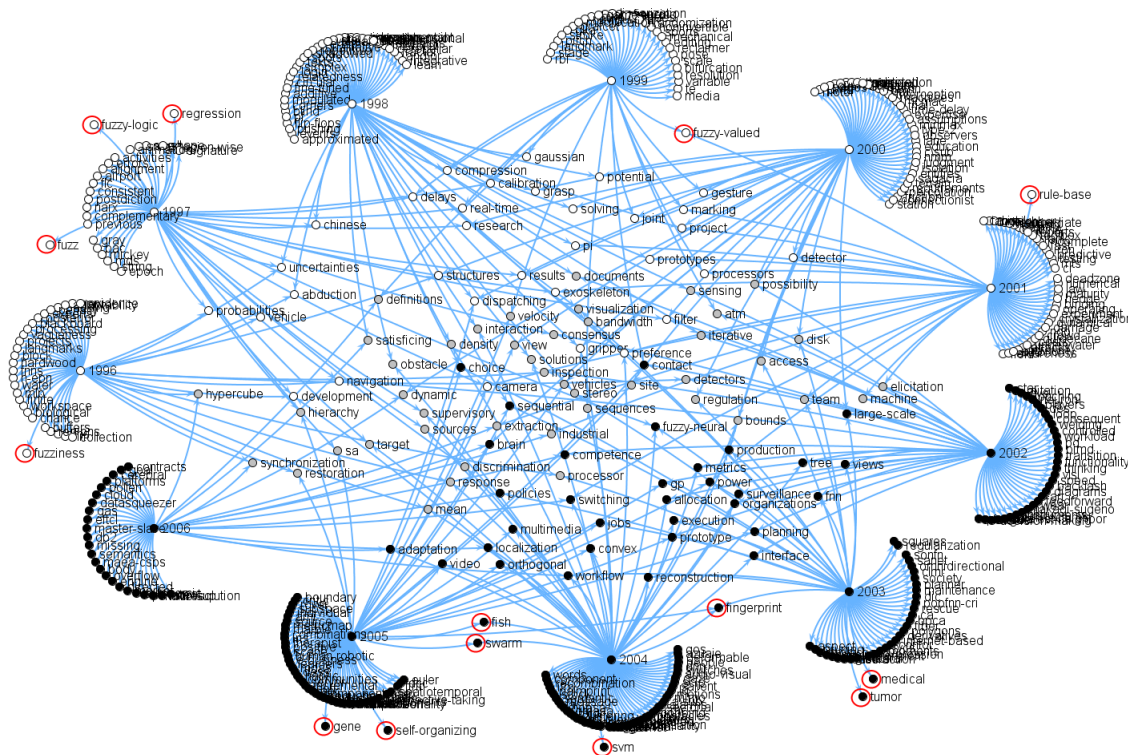
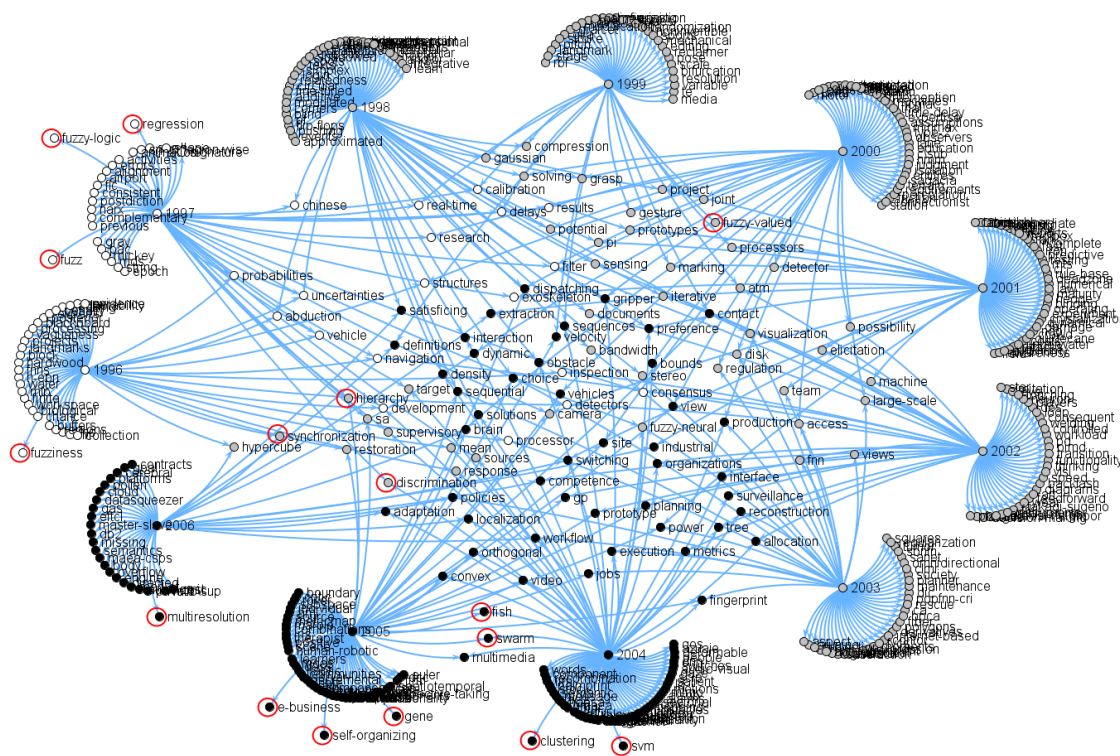Fig. 2. In this graph 1996 and 1997 are negatively primed and 2004 to 2006 are positively primed.

Fig. 3. The years were divided into older years ranging from 1996 to 2001, primed negatively, and more recent years from 2002 to 2006, which were primed positively.