Studying Evolution of a Branch of Knowledge by Constructing and Analyzing its Ontology

Pavel Makagonov,[1] Alejandro Ruiz Figueroa,[1] Alexander Gelbukh [2]

[1] Mixteca University of Technology, Huajuapan de León, Oaxaca, 69000, México.
mpp@mixteco.utm.mx, figueroa@nuyoo.utm.mx

[2] Center for Computing Research, National Polytechnic Institute, 07738, DF, Mexico
gelbukh@gelbukh.com; www.Gelbukh.com

Abstract. We propose a method for semi-automatic construction of an ontology of a given branch of science for measuring its evolution in time. The method relies on a collection of documents in the given thematic domain. We observe that the words of different levels of abstraction are located within different parts of a document: say, the title or abstract contains more general words than the body of the paper. What is more, the hierarchical structure of the documents allows us to determine the parent-child relation between words: e.g., a word that appears in the title of a paper is a candidate for a parent of the words appearing in the body of this paper; if such a relation is repeated several times, we register such a parent-child pair in our ontology. Using the papers corresponding to different years, we construct such an ontology for each year independently. Comparing such ontologies (using tree edit distance measure) for different years reveals the trends of evolution of the given branch of science.

1 Introduction

Measurement and analysis of evolution of a branch of knowledge is important for various tasks of administration and distribution of resources and human activity. For example, it is possible to predict the trends in the development of science (at least in the periods of stability) for better distribution of capital investment in scientific research [1]. In such periods development of science is based on the previous results (even though in unpredictable direction). It increases in the periods of active development of the scientific infrastructure and decreases some time before the science becomes obsolete and enters the period of stagnation. The latter is an indication of an approaching structural shift—such as a new invention—in the larger system of knowledge embracing the given branch of science, which resolves the difficulties accumulated in this subsystem.

In our previous work we have used the analysis of structured text for portraying the trends in the development of economy. We analyzed the titles of all articles published by the Free Economic Society of Russia for the period of 1983–2000 coincid-

ing with a transitional period of Russian economy [2]. The distribution of the titles in time reveals three clusters of words used in the titles of the articles:

- Words of high level of abstraction (categories of the economy field) that persist during the whole mentioned period, and
- Two different clusters of words, corresponding to the periods 1983–1989 and 1990–2000, belonging to the next, lower level of abstraction (economical processes) that have changed in Russia within the period under consideration.

The two latter clusters reveal the sharp difference between the economical state of the country in the corresponding periods. Indeed, the first period corresponds to the state-controlled economy while the second one to the free market economy.

This suggests that cluster analysis can reveal real underlying processes. In this paper we further develop this idea and apply it to the analysis of the state and development trends of a branch of science. Instead of plain clusters of terms we use a hierarchically organized terminological ontology. We illustrate the methodology on constructing an ontology on a specific thematic domain: *parallel, concurrent, distributed, and simultaneous computing*.

With this research, we introduce a novel view on ontology: not as a static collection of facts about the language and the world but as a snapshot of the state of the language and the world in a specific moment of time, which evolves as the world evolves. This leads to a novel use of (such) ontologies as a tool for describing and studying the evolution of the domain in question: we argue for that the question *what changed in the world* (*domain*)*?* is to be more precisely formulated as *what changed in its ontology?* In its turn, this thus gives a tool for qualitative and quantitative studying of the changes in a specific area of interest—which we apply in our case to scientometrics [6].

Since such "time snapshot" ontologies are nearly impossible to construct manually (indeed, their authors would inevitably reflect their own time's view, not the past moment's one), we have developed a novel method of automatic constructing such an ontology, which can be used independently whenever an ontology is to be constructed for a domain for which hierarchically-structured documents are available [7].

Our method for describing the changes in the domain through the changes in its ontology has its limitation. Indeed, changes in the domain, e.g., in science, can be reflected in three different processes:

- Appearance or disappearance of terms;
- Appearance or disappearance of relationships between terms;
- Re-conceptualizing the meaning of terms.

In this paper we mainly discuss the first effect, as the simplest and easier detectable with our methods. The second effect can also be studied by considering structural (qualitative instead of quantitative) changes in the ontology. We do not directly deal with the third effect. However, future research will show whether re-conceptualization is directly reflected in sharp changes of relationships between terms (cf. *trap a mouse* vs. *click with the mouse*) and thus is accounted for in our method.

The paper is organized as follows. Section 2 introduces the basic concepts about the type of ontology we use for our research and its interpretation for our goals. Sec-

tion 3 briefly presents the algorithm for constructing this type of ontology. Section 4 outlines the experimental results and gives an example of the constructed ontology. Section 5 provides the discussion about the use of our ontology for evaluating the changes in the science. Finally, Section 6 concludes the paper.

2 The Structure of an Ontology of Scientific Texts

By *ontology* we mean a hierarchical structure of concepts and words, which can be represented by a graph. Its root node (zero level of abstraction) identifies the thematic domain under consideration, in our case *parallel, concurrent, distributed, and simultaneous computing*. Each next level contains concepts of lower level of abstraction. The leaves of the hierarchy are the most specific terms. Note that in all algorithms we assume that the texts are stemmed [5]; in what follows for simplicity by words we actually refer to the stems.

To detect concepts and words of a given domain, we use a frequency list of common English words and a corpus of texts with the same list of words of upper level of the hierarchy: the corpus on computer science but not on parallel computing. This allows us to obtain a small but detailed ontology.

We experimented with a corpus of about 2300 articles from IEEE journals and proceedings of various conferences on parallel, distributed, concurrent, and simultaneous computing. The texts in this collection are organized hierarchically: each paper is under title of the corresponding conference or journal, the body of the paper is under its title, etc.; specifically, the hierarchy is as follows:

- Domain description: *parallel, concurrent, distributed, simultaneous computing*
  - Title of the conference or journal
    - Title of the paper
      - Abstract, introduction, conclusions, references
        - Body of the paper

The titles of the sources and of the papers are very short text segments, so we enrich each such segment with the words of the lower levels of abstraction that occur under it. The text segments of each level provide the words of the corresponding levels of abstraction. The bodies of the articles are the source of the lowest level words.

For the present study we considered three levels of abstraction:

- Titles of conferences or journals;
- Titles of papers;
- Abstracts (without titles).

We concatenated the corresponding texts dated by each year (or sometimes several years, when we had too few texts for each year): e.g., all titles of papers for 1997. This gave a relatively small group of larger files T_i (three files per year). For constructing our ontology we used vector space representation of these files with a *tf-idf*-like term weighting.

Namely, we began with constructing weighted wordlists called domain oriented dictionaries (DOD) [3]; such a list includes the domain terms along with their correspondent importance (relevance) weights for the given domain.[1] The DOD includes the words whose frequency in our domain collection is three times higher than that in the general English usage.

Using the DOD, we obtained a vector space representation (*image*) of each text T_i as a vector of frequencies of the words from the DOD in the text. These images were used to form three matrices: the *texts-by-words* matrix of the frequencies of the words from the DOD in each text; the *words-by-words* matrix of the frequencies of word co-occurrences in a text (the number of texts containing both words); and the *texts-by-texts* matrix of the numbers of DOD words in common in each pair of texts.

Figure 1 shows the text-by-word matrix (histogram) of titles. The horizontal axis corresponds to the words ordered from left to right in the chronological order of their first occurrence in a conference title or (if it does not occur in any conference title) in a paper title. The vertical axis corresponds to the documents—concatenated conference or paper titles for specific years—in the chronological order from top to down, first conference titles and then paper titles. The frequency of each word in each document is represented by the intensity of the color. The new words appearing in a given document (concatenated conference or paper titles for a year) are clearly visible as grey rectangles, and the amount of the new words is characterized by the square of this rectangle. Note that the words appearing in the conference titles have been usually seen in paper titles before: conferences are organized on what people have been discussing for a while in papers.

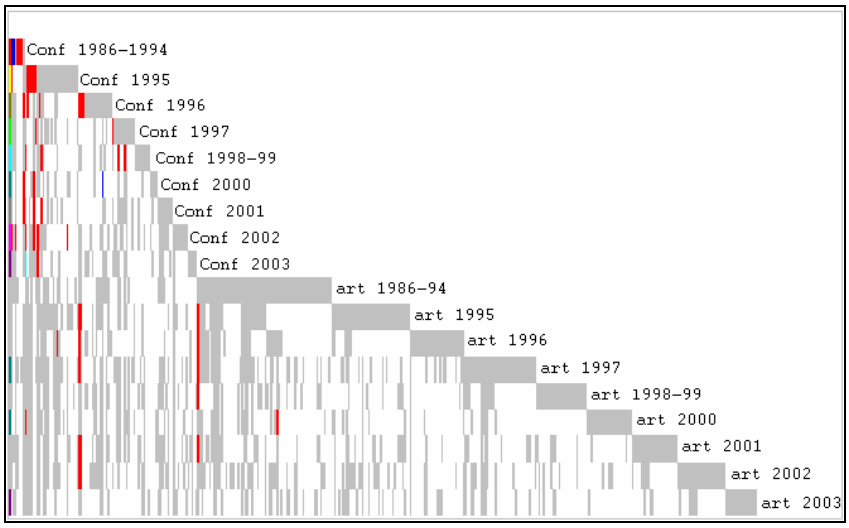

Fig. 1. The portrait of the text-by-words matrix of conference titles and article titles.

[1] For constructing the DODs we used the VHCA toolkit (*Visual Heuristic Cluster Analysis for texts*) as well as the *Administrator* and *Reader* toolkits from the IRBIS system [4].

In Figure 1 one can note two important properties of the word distribution:

- The titles of conferences contain 23% of the DOD; 13% of these words (3% of the total amount of words in the DOD) are represented also in paper titles and 77% belongs to conference titles only. This supports the idea that the 23% of the DOD belonging to conference titles are indeed words of upper level of abstraction.
- The number of new words in conference titles in 1990 to 1997 decreases (the grey rectangles become each time thinner from conferences 1995 and reach their minimum in 2000), and than a wave of new words occurs in 2001–2002 (the grey rectangles grow wider for the conferences in 2001–2002). A similar effect can be slightly noticed for paper titles. This indicates some change in the interest in main area of research, evidenced by thematic shift in conference topics.

The latter observation allowed us to subdivide the period under investigation into two stages: until 1997 and since 1998. Comparison of these two states allows determining the trends in the development of the science at their boundary. What is more, a similar analysis of a longer period would allow detecting a greater number of stages and thus more detailed development curves.

To compare the two sub-periods, we constructed two texts-by-words matrices (with the same attributes as Figure 1), one for each set of texts corresponding to the periods 1990–1997 and 1998–2003, respectively. Each such set contained the texts of different level of abstraction (conference titles, paper titles, and abstracts bodies). Using the data corresponding to each period, we constructed two corresponding ontologies. Each ontology is based on the three text-by-word matrices (conference titles, paper titles, and abstracts bodies), as described in the next section.

3 Constructing the Ontology

Consider the text-by-words matrix for the period 1998–2003 shown in Figure 2 (this matrix is transposed with respect to Figure 1, i.e., its rows correspond to terms and columns to texts). In this matrix one can distinguish a cluster ***A*** of words best represented in the conference titles. Obviously these words are also represented in the texts of lower level of abstraction, but with lower and lower intensity.

These words can belong to two different levels of abstraction:

- The root, most abstract level ***LA#1*** of our specific ontology, represented by the words *parallel*, *concurrent*, *distributed*, and *simultaneous computing* and their synonyms;
- An even greater level of abstraction: since our ontology is a part of a super-ontology of, say, *computer science*, these clusters contain also the words from this super-system, i.e., general computer science terms.

To exclude the words of the latter type, we constructed the DOD of another branch of computer science (we selected *soft computing* domain) and eliminate from our DOD those words that also belong to the DOD of this another area.

To achieve more reliable statistics, we grouped together highly co-occurring words forming so-called concepts. In our experiments, 25% of all words produced 80% of one-word concepts, the remaining non-trivial "complex clusters" containing 75% of the DOD; the number of concepts was about 20% of that of words.

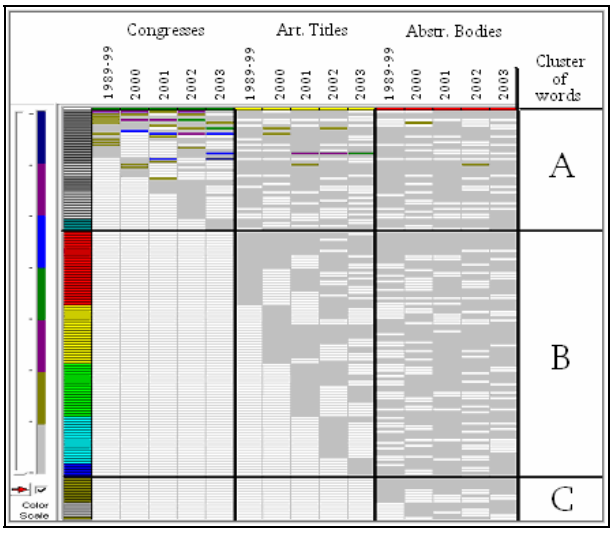

Figure 2. Distribution of clusters of words by types of documents (conference titles, paper titles, abstracts).

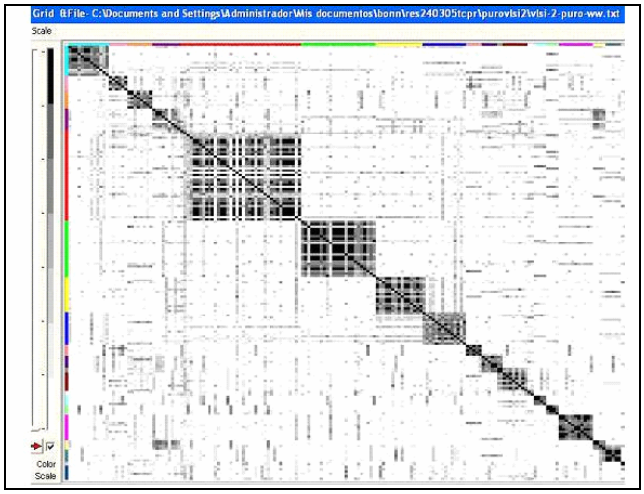

Figure 3. Clustered word-by-word matrix for the sub-base of word *VLSI* for 1998–2003.

The subdivision of the concepts (words) into the three clusters: ***A, B, C*** is the simplest form of our ontology (we call it a pre-ontology). Indeed, it is important to determine the distribution of words of clusters ***B*** and ***C*** by vertices of each sub-ontology, i.e., establish the parent-child relationships between these words and the words of ***LA#1***. For this, for every vertex (concept) of the given level ***LA#1*** we select the texts of abstracts under it containing this concept. With this we construct a sub-ontology (create its own DOD, matrices of text-by-word type, and reveal the clusters of words and concepts) for such new sub-domain, as discussed above.

Thus the process of ontology construction is recursive: the lower levels of ontology are constructed by a simple subdivision of a corresponding part of the concepts under an upper level node.

4 Experimental Results

Our initial motivation was to detect trends in computer science development over time. Accordingly, we experimented with two sets of abstracts of conference proceedings corresponding to the periods of 1990–1997 and 1998–2004. Shown below are examples from the two constructed ontologies. The words are represented by stems (not by normal form!), for example, *toleran* corresponds to *tolerance* and *tolerant*. Non-trivial concepts (synset-like ones, consisting of several words) are shown using "{ }", and non-trivial topics (closely related groups of concepts) using "[ ]". Such synset-like groups were obtained by clustering together the words with very similar co-occurrences in the texts [7].

Because of a limited space in the paper, we only show a very small excerpt from the whole ontology: the root, the first level under the root, and the nodes located in the tree under two selected nodes of the first level: the node {*analysi*, *network*, *vlsi*} and the node *toleran*. Other nodes have a similar number of sons (not shown here).

Root *parallel, concurrent, distributed, simultaneous* (computing).

1st level, below the root:

- Both periods: *{analysi, network, vlsi}, fault, orient, transaction, volum, toleran*;
- 1990–1997: *{graphic, model, securit, communicat, test}, frontier, massive, optic, real, reliabilit, reliabl*;
- 1998–2003: *{autonom, defect, discret, event, foundation, generat, grid, integrat, interact, storag, technologi, tool}, {circuit, date, evolvabl, interconnect, languag, requirement}, knowledg, object*.

2nd level, (example only for one concept) below *toleran*:

- Both periods: *{fault_toleranc}, inject, object, system*;
- 1990–1997: *{allocat, spar}, barri, board, network, critic, efficient, execut, orient, reliabl*;
- 1998–2003: *adapt, adaptat, alternat, amplifi, analog, communicat, control, controller, cost, determin, enhanc, etern, evolut, filter, flexibl, immun, java, motor, path, platform, schedul, test, tool, trigger, upgrad*.

As a gold standard for evaluation of the obtained ontology, we used a deep and elaborated structure of chapters, sections, subsections, and smaller units of good existing textbooks on the corresponding topics (i.e., the textbooks with the title corresponding to a concept from our ontology). We observed that the concepts under the given node in the tree matched well the chapters and sections of the textbooks. This is an independent proof of high quality of the automatically obtained ontology.

Of course, existing of elaborated ontology of textbooks as a gold standard does not make our work useless by providing ready results. Indeed, such a gold standard ontology extracted from a textbook can be constructed only for a period in a rather remote past for which textbook already exist. However, for the currently active research areas the textbooks do not exist yet and will appear only in the future. Actually, our ontology reflects the structure of such a future textbook (and can be used to plan such a new textbook).

The good correspondence between our ontology for a past period and the corresponding textbooks indicates that the ontology constructed with the same method for a recent period should be equally correct. However, more rigorous evaluation is a topic of our future work.

5 Stability of Ontology as a Measure of its Evolution

For determining stability of our ontology construction procedure we compared three samples of texts: two samples (of similar size) of conferences and papers for the period 1990–1997, and their union.

Accordingly, we obtained three different pre-ontologies. To measure the distance between them, we used cosine metric for distribution of words between different annual aggregated texts. Before measuring this distance, we excluded (as stopwords) some words of other systems and the super-system: numbers, toponyms, and names of events or documents (such as like *workshop*, *conference*, *transaction*).

For the first and the third (joint) samples, we obtained 62 words as vectors of their distribution by years with the mean cosine similarity 0.943 and standard deviation 0.149, and ten words that were not represented in one of the two samples (cosine is zero). However, more important characteristic of stability is the year of the first occurrence. For example, of 89 pairs of words existing in first two samplings, 13 ones have difference in the years of first occurrence. Considering difference of years of first occurrence as the distance between the words in a pair, we obtain the distance of 23 between these samples. This figure can be normalized by the maximum distance (7 years) and by the quantity of words. This approach is similar to the well-known tree edit algorithm for measuring the distance between two trees. This approach can be applied for measuring the distance between different ontology realizations, which can be used for evaluating the evolution of a given branch of science.

The ontology is a hierarchy but not a tree, thus some words can have more than one parent. However, we can measure separately the changes of both parameters (parent changing and level changing) during two given period of time.

In this case we use the criterion of evolution (total discrepancy between the ontologies of the two periods: 1990–1997 and 1998–2003) that consists of three parameters:

- appearing/disappearing of words,
- changes of the levels of words;
- changes of the parents of words.

In a process of detailing of these ontologies these figures are changing.

6 Conclusions

The hierarchical structure of technical documents is useful for automatic learning of a narrow-domain ontology from a relatively small corpus of scientific papers, as the only source of information or an additional source of evidence. The method presented here can be applied to any hierarchically structured texts, for example, HTML web pages.

Experimental results show that the constructed ontology is meaningful. Specifically, it can be used for comparative analysis of the state and development of a branch of science over different time spans. However, the most probable use of the method, as that of many other automatic ontology learning methods, is rapid prototyping of an ontology, with manual post-editing for higher-quality results.

Acknowledgements. This work was done under partial support from Mexican Government: CONACyT, SNI, SIP-IPN, COFAA-IPN. We are grateful to anonymous reviewers for useful comments.

References

1. T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 2nd edition, 1970.
2. *References Book of Proceedings of Free Economic Society of Russia*. Vol.4. (1983–2000), Moscow, Russia,.2000, pp.756.
3. P. Makagonov P., M. Alexandrov, K. Sboychakov. A toolkit for development of the domain-oriented dictionaries for structuring document flows. In: H. A. Kiers et al. (Eds.), *Data Analysis, Classification, and Related Methods*, Studies in classification, data analysis, and knowledge organization, Springer, 2000, pp. 83–88.
4. IRBIS Automated Library System, Russian National Public Library for Science and Technology; www.gpntb.ru.
5. A.Gelbukh, M. Alexandrov, S.Y.Han. Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model. *Lecture Notes in Computer Science* 3287, Springer-Verlag, 2004, p. 432–438.
6. J.A. Wagner III and R.Z. Gooding, Effects of Societal Trends on Participation Research, *Administrative Science Quarterly*, 32 (1987), pp. 241–262.

7. P. Makagonov, A. Ruiz Figueroa, K. Sboychakov, A. Gelbukh. Learning a Domain Ontology from Hierarchically Structured Texts. *Proc. of Workshop "Learning and Extending Lexical Ontologies by using Machine Learning Methods" at 22nd International Conference on Machine Learning, ICML 2005*, Bonn, Germany.