CHEMICAL ENGINEERING TRANSACTIONS Volume 17, 2009
Editor Sauro Pierucci
 ISBN 978-88-95608-01-3 ISSN 1974-9791
DOI: 10.3303/CET0917227

Retrieval of research demanded knowledge based on graphical mining

Yury, Avramenko; Andrzej, Kraslawski
Lappeenranta University of Technology
P.O. Box 20, FIN-53851 Lappeenranta, Finland; yury.avramenko@lut.fi

While most important qualitative information is contained in various diagrams and figures in form of shapes they have not been used during automated search. There has been proposed the method for retrieval of knowledge necessary for research from various sources of graphically represented data. The goal is to extract potentially interesting knowledge from a set of documents based on analysis of graphical information and next to explain the mechanism of the studied process. The method is composed of subject-driven document clustering, shape analysis, trends understanding and relevant context retrieval via semantic analysis. The proposed method is implemented in the software suite which contains source searching tool, plot comparator and semantic analyzer.
The method has been applied to find the most suitable ways to calculate the longitudinal dispersion coefficients for one branch of the Somes river, Romania, based on processing literature sources with the channel geometry characteristics of other rivers.

1. Introduction

Scientific knowledge has greatly grown in the past century. An individual scientist has to focus on or specialize in only a few scientific sub-disciplines. To advance science the researcher has to know and understand the current state of the art of the field. For this purpose the researcher has to search for any relevant information in many research reports and scientific journal which are related to the area of expertise. The amount of information sources is quite much. To help a researcher knowledge management techniques have to be used.
Knowledge management is the systematic process of finding, selecting, organizing, distilling and presenting information in a way that improves someone's comprehension in a specific area of interest. Nowadays, there are many searching Internet-based systems and many specialised electronic databases which can be used in searching relevant information using keywords. There are some problems with data search (Guha R. et al, 2002). First, the collection of information sources may contain text, images, tables of numbers and specific data-type. Processing text data requires semantic analysis. In many languages a word or phrase may have multiple meanings, so a search may result in many matches that are not on the desired topic. Another problem is incompletely defined search space as well as incomplete or noisy data. Some data may

Please cite this article as: Avramenko Y. and Kraslawski A., (2009), Retrieval of research demanded knowledge based on graphical mining, Chemical Engineering Transactions, 17, 1359-1364 DOI: 10.3303/CET0917227

be missing or is not described in the body of text whereas it represents a valuable piece of information. However, such descriptive data can be represented as images or pictograms. Finally, the amount of sources, which is relevant to topic, may be simply extremely large. Consequently, automated search engines that rely on keywords matching return too many low quality answers. Hence, after search by keywords, there are still a lot of information sources that have to be analyzed manually by the researcher.

There are a lot of graphical representation of data in chemistry and chemical engineering, such as various experimental results, represented by, for example, dependences of concentrations from time, IR spectrum, change of chemical properties with temperature etc. Those data are heavily used in chemical research to describe the process behaviour or to discover mechanism of reaction. However, whereas the graphical representation is used only for visualization, it plays key role in determination of type of phenomena of observed processes. There is a need for fast and computer-based analysis of the graphical information in order to, for example, selection of the mathematical models describing the given phenomena, selection of operational conditions basing on the graphical representation of the certain properties.

In order to fill up a gap in data analysis and to facilitate the process of finding of explanation for certain behaviour the new research approach for identification of mechanisms and trends of processes is proposed. A mining of graphical information in analogy of data mining is introduced in this work. The objective is to reuse figures in scientific articles and technical reports as well as various experimental graphical data in chemistry and chemical technology.

2. Method description

The goal is to extract interesting knowledge from a collection of information sources based on analysis of problem description containing graphical representation as a principle definition. This graphical representation could be inexplicit e.g. table with experimental data.

The method is composed of three steps:

1. Pre-selection of promising information sources which contain data related to studied problem via information retrieval techniques;
2. Qualitative comparison of graphics from information sources with the generalised shape of studied process/phenomena;
3. Retrieval of concept knowledge (e.g. mechanism description) from the source that contains graph with the most similar shape to the studied graph.

The problem should be presented in generalised form for better efficiency. Therefore, one more step is required to prepare Generalised Problem Definition (GDP) before method initiation. The result of the method is set of potentially suitable concepts which might explain behaviour described in problem definition. The entire conceptual scheme of the method is shown in Fig.1.

The method is based on determination of similarity between documents (subject analysis), curves in the graphics (shape analysis) and word meanings and terms (semantic analysis).

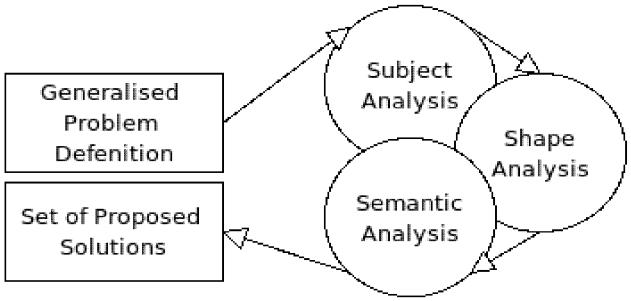

Figure 1. Outline of the method

2.1 Subject Analysis

Subject analysis is based on modified document-clustering algorithm (Zhao & Karypis, 2005). The clustering algorithm does not require any prior knowledge of the datasets in general. However, when such prior knowledge is available, clustering algorithms should also be able to benefit from it to produce more desired clustering solutions. For example, in many knowledge management applications, even though the complete taxonomy of the document collection is not available, often times domain experts can describe the major topic (clusters) that the collection covers. Moreover, they would like the clustering algorithms produce the clustering solutions that are consistent with their cognition models. Hence it is important to able to organize a document collection according to a given set of topics. It could be referred as topic-driven reasoning.

Usually, the topic in the research publications is generally represented by a set of keywords. But keywords provide rough representation of subject. There are specific words (notions) which regard to specific subject. Thus, the topics are better to represent as two subclasses of attributes: solid and amorphous identifiers.

Solid identifiers are the main keywords and action descriptions e.g. kinetics, catalyst, separation, batch reactor etc. They give a structure to the problem definition. Amorphous attributes are specific words and actions, where exact matching is not flexible in finding problem explanation in the past. Amorphous identifiers require Anchors to be defined within the construction of solid attributes. Since they are not solid it may required several terms to be identified, and each chain to an anchor should be marked by a weight. The weight shows not only degree of conformity to the solid term but also importance to current problem definition. Such identification is needed to aim fuzzy correlations in terms and avoid missing potentially useful information. For example, benzene is specific name of a compound and the sources would be relevant to the problem definition only in case of exact matching that heavily limits cluster size but if it was supplemented by such anchors as VOC, aromatic, unsaturated (term characterising current problem) then the cluster with selected documents would better cover problem subject.

2.2 Shape Analysis

Shape analysis is based on case-based reasoning (CBR) approach and general similarity concept (Avramenko & Kraslawski, 2006). The shape is presented as whether a set of

vectors or a set of polynomial curves with supplementary information about scales and presented area when available. This leads to the possibility of determining the similarity between vectors or parts of curves – they can be regarded as the problems. Since it is required only to find similar shape, therefore, no adaptation phase is needed. The main problem is the translation of the existing curve in the current source (case) into dimensionless shape to be comparable with those in GPD.
The shape is remembered as a set of proportions of the curve regardless of the absolute values and the scale of the graphics. Only the most characteristic part of the curve may be generalized as the researcher could be interesting only in certain behaviour. The closest match of subset of the curve under consideration and the generalized shape indicates a possibly similar shape.

2.3 Semantic Analysis

The analysis is purposed to find the knowledge in form of text which is most relevant to the shape. Text mining techniques are used for it (Beliaev & Kraslawski, 2005). Particularly related text could be found by decomposition of the sentences according to the typical structure. The part of text containing the specific identifiers could be retrieved and content analysis is done to understand the text (Yeh et al., 2005). As a result, an abstract of the retrieved knowledge is suggested as a description for the seeking solution or explanation of the behaviour or phenomenon.

Implementation

The described approach is realized in the software research suite. It composes of four components: getting search tool, the document clusterizer, the plot comparator and the content retriever. They implement corresponding phases of the method: pre-selection of documents (first two components), graph recognition and analysis, content retrieval. The preliminary stage of preparation of the GPD is not implemented as a tool. The solids and amorphous identifiers are assigned manually during the development of problem definition. The user should also prepare the generalised shape with most characteristic parts before starting the analysis.

3.1 Document search and clustering

The function of first tool is,information retrieval technique, to get list of sources which correspond to problem subject. The process starts with ordinal keyword search based on two types of identifiers in the various on-line search systems (e.g. Science Direct). The generated list is retrieved and prepared for analysis by means of getting search tool.
The retrieved sources are grouped as the clusters utilizing the topic-driven clustering method, which is implemented in the document clusterizer. The clusters are built accordingly to the similarity of notions. The parameters of the cluster (as centroid notion, similarity functions) can be varied during interventions in the configuration file of the tool. Next, the clusters are selected according to the problem descriptors, combined, refined, and the preliminary list of the sources is generated.

3.2 Graphics comparison

Second module is purposed to detect graphical site in the source, to recognise borders, axes and grid, to read the shape of curve or curves and finally to compare with the shape from generalised problem description.

3.3 Retrieval of content

Last module is a side-tool for semantic analysis of the text and is supposed to retrieve the content which is related to graphic explanation or description..The tool should be able to detect the area of the text corresponding to the retrieved graphical information, understand the meaning of the content and propose the solid concept relevant to the problem under consideration. For these purposes, Text Miner (Beliaev & Kraslawski, 2005) could be used. The search and testing of the most appropriate tool with combination of the content similarity technique is subject of current research.
The overall algorithm of the tool is shown in Fig. 2.

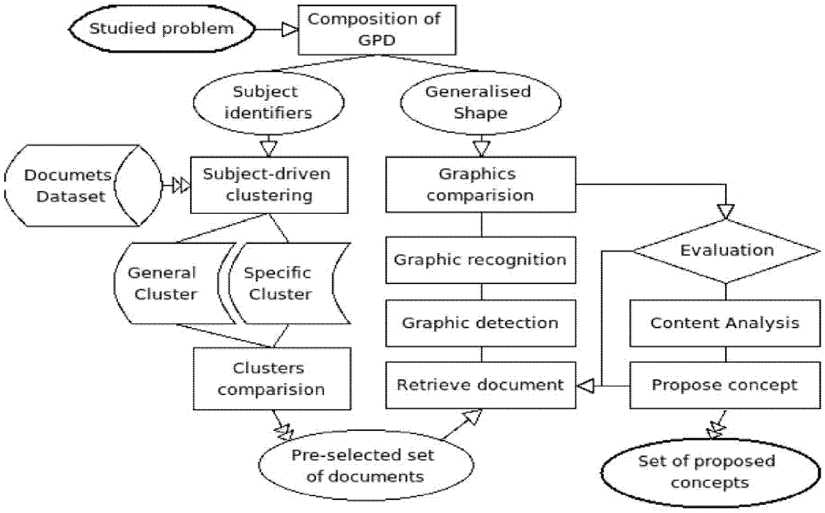

Figure 2. Scheme of algorithm for retrieval of research demanded knowledge

4. Applicaions

The developed software has been used to detect the mechanism of production of biochemical compound based on comparison of the kinetic experimental data and for to calculate the longitudinal dispersion coefficients for a branch of the river.

4.1 Detection of mechanism of biochemical reaction

The objective is to detect the mechanism of production of chemical compound using the experimental data. It is known that the compound is produced during growth of microorganisms. Thus the subject is defined as microbiology and fermentation. The concentration profile of compound is observed as shown in Fig.3a. The problem data is represented as dimensionless plot which depicts general trend in concentration. There have been searched appropriate graphical representations of kinetics that could correspond to the observed data. The most similar concentration curve has been identified for penicillin according to given information source (Fig.3b). There can be suggested the same mechanism of production for investigating compound.

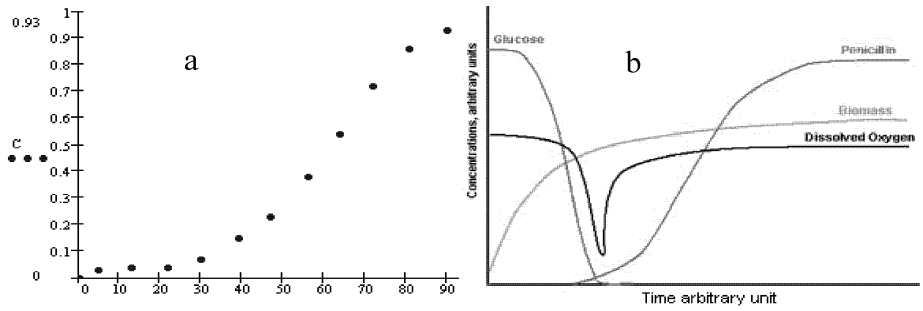

Figure. 3. Concentration curve of tested compound (a) and concentration profiles of the most similar source (b)

4.2 Identification of methods for calculation of dispersion coefficients in the rivers

The objective is to find relevant description (empirical formulas, theoretical models and estimations) of the dispersion coefficient determination in those river channels that have similar geometry to the tested river branch.

The empirical formulas rely on the physical characteristics of the river channel, therefore such characteristics as longitudinal and transversal profiles of the channels are used as comparison criteria. The experimental data was collected in year period along the reaches of the river. The measurements include channel width, depth and bed slope.

After determination of the list of the sources correspond to the subject, the procedure of shape comparison was divided into two processes: based on longitudinal profiles (L-profiles) and based on transversal profiles (T-profiles). The Plot Comparator found three sources with similar T-profiles and 2 with similar L-profiles. The documents containing similar profiles were selected for the concept retrieval.

The text mining tool was applied to retrieve the content relevant to the dispersion coefficient calculation. The retrieved knowledge about dispersion coefficients offers several techniques for their estimation, which are applicable for the different river geometries. It is important to have the detailed data about transversal profile to use these empirical formulas or the models.

References

Avramenko, Y., Kraslawski, A., 2006, Similarity concept for case-based design in process engineering, Computers & Chemical Engineering, 30, Issue 3, 548-557.

Beliaev, S. and Kraslawski, A., 2005, Text mining for identification of new products and multifunctional materials, In Proc. 7th World Congress of Chemical Engineering, Glasgow, Scotland.

Guha, R., McCool, R., Miller, E., 2003, Semantic Search, In Proceedings of the 12th international conference on World Wide Web, 700-709.

Yeh, Jen-Yuan, Ke, Hao-Ren, Yang, Wei-Pang, Meng, I-Heng, 2005, Text summarization using a trainable summarizer and latent semantic analysis, Information Processing and Management 41, 75-95.

Zhao, Y., Karypis, G., 2005, Topic-driven clustering for Document Datasets, In Proceedings of SIAM International Conference on Data Mining, 358-369.