17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

Mining of Graphics for Identification of Mechanisms and Trends of Processes

Yuri Avramenko, Andrzej Kraslawski

Department of Chemical Technology, Lappeenranta University of Technology, P.O. Box 20, FIN-53851 Lappeenranta, Finland; Andrzej.Kraslawski@lut.fi

Abstract

The paper describes a method for identification of mechanisms and process trends based on combination of subject-driven document clustering, shape analysis, trends understanding and relevant context retrieval via semantic analysis. The goal is to extract potentially interesting knowledge from a set of technical information based on analysis of graphical information in order to find explanation for a specific process behavior.

Keywords: shape comparison, similarity measurement, concept retrieval

1. Introduction

The over-supply of data, information and knowledge is a well-known problem in R&D activities. One of the approaches aimed at limiting the negative impact of the amount of information is its reuse. There are two major possibilities of information reuse. The first one is a search for the solution of a new problem basing on the results obtained for the past, similar cases. The second approach is knowledge discovery consisting in the compilation of the information from the various sources. In both cases, information reuse of the unstructured data, e.g. scientific articles, technical reports and patent descriptions, has been limited to text mining utilizing mostly syntax analysis and keywords searching. Unfortunately, the most information-rich sources, diagrams and figures, have not been used in information reuse. While most important qualitative information is contained in various charts in form of shapes. Even if during

routine keywords search the sources with such charts were selected then an engineer may encounter a problem to determine the shape similarity between considered problem and historical data. In order to fill a gap in data analysis and to facilitate the process of finding explanation for certain behavior the novel approach for identification of mechanisms and trends of processes is proposed.
A mining of graphical information in analogy to data mining is introduced in this paper. Its objective is to reuse figures in scientific articles and reports.

2. General Outline

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 the 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 one.

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 a set of potentially acceptable concepts which may 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). Similarity measurement is based on generic principles of General Similarity Concept which is described shortly together with other techniques in next section.

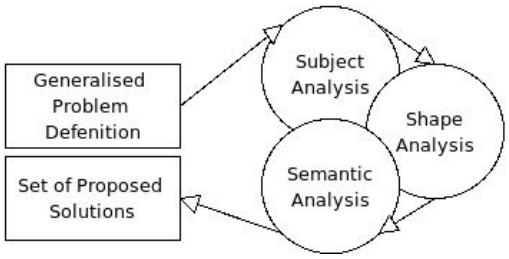

Figure 1. General outline of the method

3. Method description

3.1. Problem Representation

The problem description must provide information about subject and shape characteristics. A subject in a research publication is generally represented by a set of keywords. But keywords provide rough representation of subject. Thus, complete problem description is represented as three classes of attributes: solid identifier, amorphous identifiers, and generalised shape.
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 terms are specific words and actions, where exact matching is not flexible in finding problem explanation. Amorphous identifiers require the anchors to be defined within the construction of solid attributes. Since they are not solid it may require 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. It serves 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 suit the problem only in case of exact matching that heavily limits cluster size but if it was supplemented by such anchors as VOC, aromatic, unsaturated then the cluster with documents would better cover problem subject. The shape is remembered as a set of proportions of the curve regardless to absolute values and scale of the graphics. Only most characteristic part of curve might be generalized to be more sensitive.

3.2. General Similarity Concept

The concept gives a way of determination of similarity between documents, shapes, sets, vectors and numbers based on few basic definitions.
First, any piece of information is represented as an entity. An entity description includes: F- a list of features, R - a set of relations between them, and V - a set of feature values. Thus, the entity E is defined as follows:

$$E = < F, V, R >, \quad (1)$$

The representation may be extended by including attributes of features which usually reflect a degree of importance of corresponding feature. In such a case the representation of an entity is supplemented with W – set of weights.
Further, generic definition of similarity is introduced. The *degree of similarity* is quantitative measure defined by a ratio of number of matched features of two entities to overall number of features. When features have spectre of values exact match is replaced to fuzzy match (interval from 0 to 1).

Applying only such definitions the degree of similarity between two entities E_1 and E_2 of same weighted representation is calculated as follows:

$$SIM(E_1, E_2) = \sum_{i=1}^{k} w_i \cdot sim_i \Big/ \sum_{i=1}^{k} w_i \qquad (2)$$

where k is number of features in the structure, and w_i – weights of importance.
The degree of similarity is denoted as complementary to degree of difference, and therefore the similarity between the feature values is defined as

$$sim(a,b) = 1 - d(a,b) \qquad (3)$$

where $d(a,b) \in [0,1]$ is a difference (distance) of two values a and b. If the features are completely different then d equals 1. The difference measurements are dependent on data type of values. They are derived for specific data type from basic definition of difference. The examples of measurements for basic type of data are shown in table 1.

Table 1. Examples of basic difference functions

Type	Measurement		Type	Measurement
Numeric	$d = \frac{\lvert a-b \rvert}{range}$	*range* – a range of values of variable *a* and *b*.	Sets	$d = 1 - \frac{\lvert a \cap b \rvert}{\lvert a \cup b \rvert}$
Vectors	$d = \frac{\lvert \vec{a} - \vec{b} \rvert}{\left\lvert \sum_{i=1}^{n} \vec{e}_i \right\rvert}$, vectors *a* and *b* are normalized	$\vec{e}_1 = (1;0;\ldots;0)$ ⋮ $\vec{e}_n = (0;0;\ldots;1)$	Logical	$d = \begin{cases} 1, a \neq b \\ 0, a = b \end{cases}$

3.3. Document Clustering

The documents in the dataset are represented as sets of terms. The objective is to organize dataset according to a given set of subjects describing the problem.
The step utilises classical vector space model of information retrieval and topic-driven clustering method [1]. The set of problem's terms (identifiers) are divided in two subsets – S (solid terms) and A (anchors of amorphous terms).
Similarity between a document and the problem subject – both are represented as sets – is based on the difference measurement for sets (see table 1).
The documents are organized into three clusters: relevant to general topic (represented by S), relevant to specific topic (mostly A set) and not relevant to problem at all. Sets S and A create centroid sets for own clusters. Each amorphous identifier is represented as an entity because of complex structure (includes relation weights). The similarity is determined as between entities.

The pairwise similarities of two clusters are then determined to indicate the most correlated documents. These documents build preliminary set of sources.

3.4. Graphics comparison

A recognised graphic from an information source is represented as a set of vectors. The curve is decomposed of short lines which are translated into vectors. The set of vectors is compared with the generalised shape which is also represented as set of vectors. The closest match of subset of the curve under consideration and the generalized shape indicates promising shape.

3.5. Concept retrieval

The text from selected information sources is "read" to extract a set of semantic features. Contextual-usage meaning of words are retrieved via latent semantic analysis [2]. The semantic similarity of terms (similarity of meaning of words or sets of words to each other) is considered to avoid strict matching. The subset of features that is semantically most similar to set of problem's identifiers is used to create summary of concept – a proposed solution to the problem.

4. Algorithm design and implementation

Complete algorithm scheme is represented in the Fig. 2.
The method is being realized in the software package: source searching tool, plot comparator and semantic analyzer. They implement different phases of analysis: subject, shape and semantic analysis correspondingly.
The first tool utilises subject-driven clustering to select documents relevant to problem subject. Second module is purposed to detect graphical site in the source, to recognize border, axes and grid, to read the shape of curve or curves and finally to compare with the shape from GPD. Last module which is not implemented yet is supposed to text analysis to find content explaining detected figure. The Text Miner [3] software is promising to perform this task.

5. Illustration of method

The method has been tested using the data generated by the authors.
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) – the degree of similarity is 0.94. It could be suggested the same mechanism of production for investigating compound.

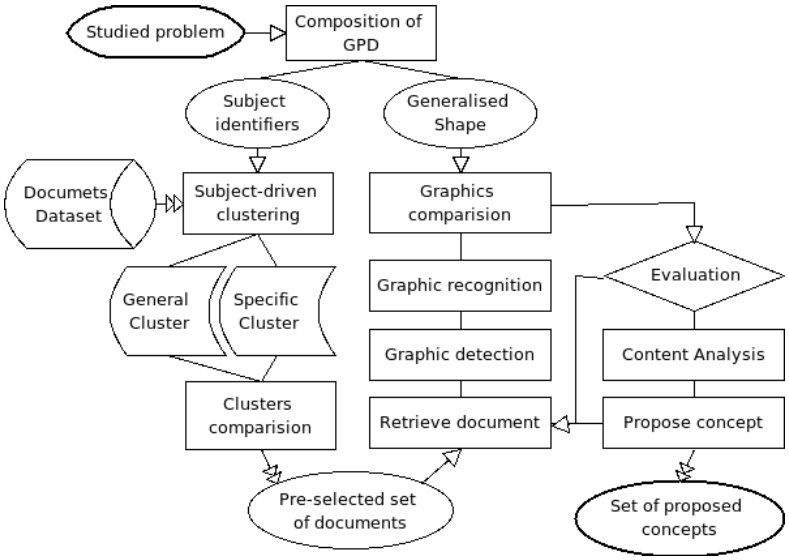

Figure 2. Algorithm scheme

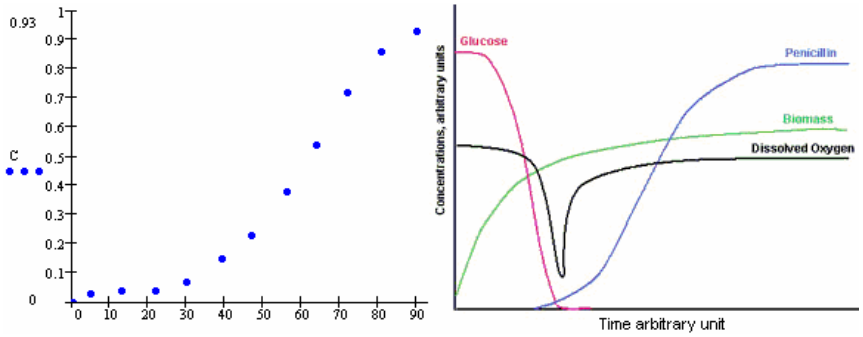

Figure. 3. Concentration curve of tested compound (a) and concentration profiles of the most similar source (b)

References

1. Y. Zhao, G. Karypis, SIAM International Conference on Data Mining, pp. 358-369, 2005
2. J.-Y.Yeh et al., Information Processing and Management 41, pp. 75-95, 2005
3. S. Beliaev, A. Kraslawski, 7th World Congress of Chemical Engineering, Glasgow, 2005