An Unsupervised Model for Rule-Based Similarity Learning from Textual Data*

Andrzej Janusz[1], Dominik Ślęzak[1,2]

[1] Faculty of Mathematics, Informatics and Mechanics, The University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
[2] Infobright Inc.
ul. Krzywickiego 34 lok. 219, 02-078 Warsaw, Poland
andrzejanusz@gmail.com,slezak@infobright.com

Abstract. This paper presents a preliminary research on construction of a new unsupervised model for learning a semantic similarity measure from text corpora. Two main components of the model are a semantic interpreter of texts and a rule-based similarity function. The first one associates particular documents with concepts defined in a knowledge base which corresponds to the topics covered by the corpus. It shifts the representation of a meaning of the texts from words that can be ambiguous to concepts with predefined semantics. With this new representation, the similarity function is derived from data using a modification of the Dynamic Rule-Based Similarity (DRBS) model, which is adjusted to the unsupervised case. This adjustment is based on a novel notion of information bireducts. This extension of classical information reducts is used in order to find diverse sets of reference documents that determine different aspects of the similarity. The paper explains a general idea of the approach and gives some implementation guidelines.

1 Introduction

Selection of appropriate similarity measure for a given task has always been a challenge to researchers in fields related to data mining and information retrieval. Although numerous similarity functions have been proposed, none of them is general enough and different applications require different functions to be used. In practice, a similarity measure which is reasonable for a given task is chosen by experts but, due to the complexity of the similarity evaluation problem, this choice is rarely optimal.

To overcome this issue, some techniques for learning the similarity from data have been developed ([1], [2], [3], [4]). Instead of relying on globally defined measures, those methods try to discover proper local similarity functions and

* The authors are supported by the grant N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland and by the National Centre for Research and Development (NCBiR) under the grant SP/I/1/77065/10 by the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

CONCURRENCY, SPECIFICATION AND PROGRAMMING
M. Szczuka et al. (eds.): Proceedings of the international workshop CS&P 2011
September 28-30, Pułtusk, Poland, pp. 229-237

aggregate them in a way that reflects dependencies between objects from particular dataset. For this purpose, they often make use of basic properties of the similarity relation in a given context. For example, the Dynamic Rule-Based Similarity (DRBS) modela [1] utilizes the fact that two objects from different decision classes can not be similar. Other models restrain the search space to some classes of the similarity functions (e.g. distance-based measures [2], [3], [4]).

One application of similarity models is the clustering task. Unlike in the supervised classification case where the context for the similarity is defined by a decision attribute, the clustering requires evaluation of resemblance in a general setting. Depending on a domain of objects, the context for similarity in the clustering task can be, e.g., a "general appearance" of physical objects or a "meaning" of texts. Such semantic ambiguity makes the selection of appropriate similarity measure even more challenging. Currently, only few similarity learning models are able to cope with it.

Classical clustering algorithms utilize distance measures to compute dissimilarities between objects. By doing so, they enforce some potentially undesired properties on the resulting similarity relation and make it less reflecting the human perception of similar objects. This drawback is especially important for grouping semantically related documents within text corpora.

Many methods utilizing different similarity measures were developed for the purpose of the textual data clustering. The most popular ones are modifications of the standard k-means algorithm, which use some spheric distance measures such as the *cosine distance* ([5], [6]). Those approaches are usually based on a *bag-of-words* representation of documents, in which each text is represented by a vector of weights assigned to unique terms that it contains.

There have been many attempts to extend the bag-of-words representation in order to more accurately capture semantics of texts. It has been done by, for example, inclusion of tags derived from linguistic analysis of natural language or considering n-grams of words [5]. A different method has also been developed within the rough set theory. In the Tolerance Rough Set Model the bag-of-words representation is extended by considering the upper approximation of texts ([7], [8]). Other approaches, such as the Latent Semantic Analysis [9], tried to directly model concepts related to the texts. In [10] the Explicit Semantic Analysis model (ESA) is described, in which documents are represented by their associations with concepts explicitly defined in an external knowledge base. Such a representation of texts will be called the *bag-of-concepts*.

The unsupervised similarity learning model proposed in this paper makes use of ESA. It treats the associated concepts as basic features, using which the notion of similarity between documents can be expressed. The DRBS model is adapted to the unsupervised case in order to learn aggregations of the basic features. The resulting higher-level feature sets can be regarded as different aspects of similarity, which are meaningful for a given text corpus, considering the available knowledge base. A brief overview of the proposed model is presented in the further sections. It is followed by a description of experiments which are being designed in order to evaluate usefulness of this approach.

2 Preliminaries

The main idea of the model that we propose is to utilize a variant of the psychologically plausible Tversky's contrast model of similarity (see [11]) for evaluation of resemblance between semantic representations of documents. Features that are taken for the contrast model are defined and aggregated by an analogy to the DRBS approach [1] – they can be derived from the text corpus using ESA (see [10], [12]) and information bireducts [13]. In this section we explain some basic notions that we use in construction of the model.

2.1 Information bireducts

The problem of learning a similarity relation from data involves working on imprecise concepts and it may be well-handled in a framework provided by the rough set theory [14]. In this setting available objects are often described within an *information system* $\mathbb{A} = (U, A)$, where U is a set of objects and A is a set of their attributes.

An information system may be seen as a tabular representation of knowledge about a considered universe. We can also represent in this way a text corpus, e.g., by taking all unique terms from the corpus as attributes whose values correspond to relevance of a given term to a selected text (see [5], [6]). More information on the text representation that we use is given in Section 2.3.

Two of the key concepts for the rough set theory are a discernibility relation and an information reduct. We say that two objects $u_1, u_2 \in U$ can be discerned in an information system $\mathbb{A} = (U, A)$, if and only if there exists at least one $a \in A$ for which values $a(u_1)$ and $a(u_2)$ are sufficiently different (e.g.: u_1 and u_2 are discernible in a classical sense when $a(u_1) \neq a(u_2)$). Moreover, we say that $B \subseteq A$ is an information reduct for $\mathbb{A} = (U, A)$, if and only if it is an irreducible subset of attributes such that each pair of objects which is discerned by A, is also discerned by B.

In [13] we have extended the notion of an information reduct to an information bireduct, which is a pair (B, X) consisting of a subset of attributes B and a subset of objects X, defined in the following way:

Definition 1. *Let $\mathbb{A} = (U, A)$ be an information system. A pair (B, X), where $B \subseteq A$ and $X \subseteq U$, is called an information bireduct, if and only if B discerns all pairs of objects in X, and the following properties hold:*

1. *There is no proper subset $C \subsetneq B$ such that C discerns all pairs in X;*
2. *There is no proper superset $Y \supsetneq X$ such that B discerns all pairs in Y.*

The information bireducts describe non-extendable subsets of objects that are discernible using irreducible subsets of attributes. They may be regarded as corresponding to the most irregular, informative areas of data. We use this property in our unsupervised similarity learning model to minimize the description length and maximize diversity of reference object sets.

2.2 The Tversky's Model

In 1977, Amos Tversky, influenced by the results of his experiments on human perception of the similarity, came up with the *contrast model* of similarity ([11]). He argued that the distance-based approaches are not appropriate for modeling similarity relation due to constrains imposed by the mathematical features of the distance measures. He also noticed that the similarity between objects depends not only on their common features but also on the features that are considered distinct. Such features may be interpreted as arguments for or against the similarity. He proposed the following formula to evaluate the similarity of the compared stimuli:

$$Sim(x,y) = a\,|X \cap Y| - b\,|Y \setminus X| - c\,|X \setminus Y|\,, \tag{1}$$

where X and Y are the sets of binary features of the instances x, y and the constants a, b, c are the parameters. Depending on the values of a, b, c the contrast model may have different characteristics, e.g., for $b \neq c$ the model is not symmetric. Using that model Tversky was able to create similarity rankings of simple geometrical objects which were consistent with evaluations made by humans.

However, the contrast model has a major drawback. It might be considered as impractical because there is no efficient way of finding optimal values of the parameters and it is very difficult to define the set of binary features important in the specific context of the similarity. This set may consist of some higher-level characteristics of instances which usually are not included in the information systems as attributes. This problem is even more significant when the contrast model is applied for semantic comparison of text documents.

Usually, texts are represented by associations with words which they contain. Due to a large number of possible words and their ambiguity it is not clear whether a particular word is truly relevant to express the semantic of a document. Even if two documents share the same word, its meaning can be different and, as a consequence, such a word should not be regarded as a common feature. On the other hand, different words can be related, expressing the same semantic entity. For those reasons semantic similarity of texts should be considered in terms of well-defined concepts, not just words.

2.3 Explicit Semantic Analysis

Any text document can be represented by predefined concepts which are related to the information that it carries (its semantic). One method for constructing such representations is Explicit Semantic Analysis ([10], [12]). In this approach, natural language definitions of concepts from an external knowledge base, such as an encyclopedia or an ontology, are matched against documents to find the best associations. A scope of the knowledge base may be general (like in the case of Wikipedia) or it may be focused on a domain related to the investigated text

corpus, e.g. Medical Subject Headings (MeSH)[3] (see [15]). The knowledge base may contain additional information on relations between concepts, which can be utilized during computation of the "concept-document" association levels. Otherwise it is regarded as a regular collection of texts, with each concept definition treated as a separate document.

The associations between concepts from a knowledge base and documents from the corpus are treated as indicators of their relatedness. They are computed two-fold. First, after the initial processing (stemming, stop words removal, identification of terms), the corpus and the concept definitions are converted to the *bag-of-words* representation. Each of the unique terms in the texts is given a weight expressing its association strength.

Assume that after the initial processing of a corpus consisting of M documents, $D = \{T_1, \ldots, T_M\}$, there have been identified N unique terms (e.g. words, stems, N-grams) $w_1, \ldots, w_N$. Any text T_j in the corpus D can be represented by a vector $\langle v_1, \ldots, v_N \rangle \in \mathbb{R}_+^N$, where each coordinate v_i expresses a value of some relatedness measure for i-th term in vocabulary (w_i), relative to this document. The most common measure used to calculate v_i is the *tf-idf* (term frequency-inverse document frequency) index (see [5]) defined as:

$$v_i = tf_{i,j} \times idf_i = \frac{n_{i,j}}{\sum_{k=1}^{N} n_{k,j}} \times \log\left(\frac{M}{|\{j : n_{i,j} \neq 0\}|}\right), \tag{2}$$

where $n_{i,j}$ is the number of occurrences of the term w_i in the document T_j.

In the second step, the bag-of-words representation of concept definitions is transformed into an inverted index which maps words into lists of K concepts described in a knowledge base, $c_1, \ldots, c_K$. The inverted index is then used to perform a semantic interpretation of documents from the corpus. For each text, the semantic interpreter iterates over words that it contains, retrieves corresponding entries from the inverted index and merges them into a weighted vector of concepts that represents the given text.

Let $T = \langle w_i \rangle_{i=1}^{N}$ be input text, and let $\langle v_1, \ldots, v_N \rangle$ be its tf-idf vector, where v_i is the weight of the term w_i. Let k_{ij} be an inverted index entry for w_i, where k_{ij} quantifies the strength of association of the term w_i with a knowledge base concept c_j, $j \in \{1, ..., K\}$. The new vector representation of T is calculated as:

$$\langle \sum_{i:w_i \in T} v_i k_{ij} \rangle_{j=1}^{K}. \tag{3}$$

This new vector will be called a *bag-of-concepts* representation of a text. If the utilized knowledge base contains additional information on semantic dependencies between the concepts, this knowledge can be used to further adjust vector (3). However, particular methods of doing that are not in the scope of this research.

[3] MeSH is a controlled vocabulary and thesaurus created and maintained by the United States National Library of Medicine. It is used to facilitate searching in life sciences related article databases.

The bag-of-concepts representation makes it possible to examine relations between concepts and documents as well as to identify and filter key concepts for given documents in a corpus. The concepts with the strongest association levels with texts can be regarded as their semantic features. This fact will be used in the proposed approach to instantiate the Tversky's similarity model. The whole process will be further optimized by utilization of information bireducts (see Section 2.1) in order to unbias the model with regard to hidden relations between the concepts.

3 The Model Design

In this section we show how the intuition behind the Tversky's model can be used for evaluation of semantic similarity of scientific articles. The construction of the model starts with assigning concepts from a chosen knowledge base to a training corpus of documents. This can be done in an automatic fashion with the use of the ESA method, as explained in Section 2.3.

The key concepts assigned to the documents can be treated as binary features and therefore, are suitable to use with the contrast model of similarity. However, a direct application of this model would not take into consideration data-based relations between concepts from the knowledge base. The problem of finding appropriate values of parameters would also remain unsolved. In the presented model, a rough set approach is used to overcome those issues.

Let F be a set of all possible key features of texts from a corpus D and F_T be a set of key (the most important) concepts related to the document T, $F = \bigcup_{T \in D} F_T$. If we treat concepts from F as features, we can construct an information system $S = (D, F)$. In order to find out which combinations of concepts comprise the informative aspects of similarity, we could compute information reducts of S. However, to ensure stability of the model and to limit its bias toward common objects and concepts of negligible importance, we suggest working with the information bireducts (see Section 2.1). For more details regarding theoretical foundations of bireducts along with some practical algorithmic solutions for their computation one may refer to [13].

For each bireduct $BR = (B, X)$, $B \subseteq F$, $X \subseteq D$, we can define a commonality relation in D with regard to BR. One example of such a relation can be $\tau|_{BR}$ which is defined as follows:

$$(T_1, T_2) \in \tau|_{BR} \iff T_2 \in X \wedge \left|F_{T_1|_{BR}} \cap F_{T_2|_{BR}}\right| \geq p, \tag{4}$$

where $p > 0$, $T_1, T_2 \in D$ and $F_{T|_{BR}} \subseteq B$. Intuitively, two documents are in the commonality relation $\tau|_{BR}$ if and only if one of them is covered by the bireduct BR and they have at least p common concepts. We will denote the commonality class of a document T in with regard to BR as $I(BR, T)$.

Having defined the $I(BR, T)$ we can finally utilize the idea of the Tversky's contrast model in order to compute the similarity between any two documents from D in a context corresponding to BR. It can be done using the following

formula:

$$Sim_{BR}(T_1, T_2) = \frac{|I(BR, T_1) \cap I(BR, T_2)|}{|X|} - \frac{\left|\left(X \setminus I(BR, T_1)\right) \cap I(BR, T_2)\right|}{|X|}. \tag{5}$$

Since each information bireduct in our setting is a non-extendable subset of documents coupled with an irreducible subset of features that discern them, it carries maximum information on a diverse set of reference documents. Due to this property, the utilization of bireducts nullifies the undesired effect that common objects (or usual features) would have on the sizes of commonality classes. We also believe that, by analogy to the initial experiments with *decision bireducts* (see [13]), an ensemble of the information bireducts will cover much broader aspects of data than the regular information reducts and will contribute to better performance of the model.

The similarities in many different aspects (which correspond to different bireducts extracted from data) can be easily aggregated by taking the following average:

$$Sim(T_1, T_2) = \frac{\sum_{BR} Sim_{BR_i}(T_1, T_2)}{\#extracted\ bireducts}. \tag{6}$$

Design of such a similarity function is computationally feasible and does not require tuning of unintuitive parameters. Additionally, it guarantees that the resulting similarity model keeps the psychologically plausible properties of the contrast model.

4 Directions for the Future

Usefulness of the proposed model in practical application still needs to be evaluated. The best way of doing this is to experimentally compare the unsupervised DRBS with other similarity models on real-life data. At the moment, we are designing such an experiment. A quality of an unsupervised similarity model can be evaluated by investigation of clustering results performed by some common algorithms. In our initial experiments we would like to measure how well our model is able to arrange documents into semantically homogeneous groups. The first dataset we would like to use is the corpus utilized in [12]. It contains scientific papers related to research on rough set theory and since the notion of rough sets is close to us, we will be able to assess the clustering results and manually compare different similarity models. We also have plans for preparing an experiment on medical articles from the PubMed repository [16] in which we would like to utilize MeSH [15] as the domain ontology. In this test we are going to estimate performance of the proposed similarity learning model using categorizations made by experts from PubMed Central.

Another research direction involves further investigation of the information bireducts and their properties. For example, we would like to verify whether there exists any correlation between properties of the information bireduct ensembles which are imposed by their generation process [13] and performance of

the unsupervised DRBS model for textual data. Discovery of such dependencies may help to improve the resulting similarity relation and speed up its induction. It would also be interesting to compare the unsupervised DRBS models constructed using the information bireducts and the classical information reducts. This comparison would help to better evaluate practical benefits from using the bireducts.

Finally, using the proposed model, it is possible to experiment with different parameter settings for commonality relations and information bireduct ensembles. In order for the model to be applicable to large article repositories, it is necessary to develop fast heuristic algorithms for computing bireducts and finding reasonable parameter values.

References

1. Janusz, A.: Utilization of dynamic reducts to improve performance of the rule-based similarity model for highly-dimensional data. In: Proc. of Int. Conf. on Web Intelligence and International Conference on Intelligent Agent Technology (WIIAT) – Workshops, IEEE (2010) 432–435
2. Nguyen, S.H.T.: Regularity analysis and its applications in data mining. PhD thesis, Warsaw University, Faculty of Mathematics, Informatics and Mechanics (1999) Part II: Relational Patterns.
3. Stahl, A., Gabel, T.: Using evolution programs to learn local similarity measures. In: In Proceedings of the Fifth International Conference on Case-Based Reasoning, Springer (2003) 537–551
4. Xiong, H., Chen, X.w.: Kernel-based distance metric learning for microarray data classification. BMC Bioinformatics **7**(1) (2006) 299
5. Feldman, R., Sanger, J., eds.: The Text Mining Handbook. Cambridge University Press (2007)
6. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley, Boston (2006)
7. Ho, T.B., Nguyen, N.B.: Nonhierarchical document clustering based on a tolerance rough set model. International Journal of Intelligent Systems **17** (2002) 199–212
8. Ngo, C.L., Nguyen, H.S.: A tolerance rough set approach to clustering web search results. In Boulicaut, J.F., Esposito, F., Giannotti, F., Pedreschi, D., eds.: Knowledge Discovery in Databases: PKDD 2004. Volume 3202 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2004) 515–517
9. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41**(6) (1990) 391–407
10. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. (2007) 6–12
11. Tversky, A.: Features of similarity. Psychological Review **84** (1977) 327–352
12. Szczuka, M., Janusz, A., Herba, K.: Clustering of rough set related documents with use of knowledge from dbpedia. In: Proc. of Int. Conf. on Rough Sets and Knowledge Technology (RSKT). LNAI, Springer Berlin/Heidelberg (2011)
13. Ślęzak, D., Janusz, A.: Ensembles of bireducts: Towards robust classification and simple representation. In: Proc. of Int. Conf. on Future Generation of Information Technology (FGIT). LNCS (2011)

14. Pawlak, Z.: Information systems, theoretical foundations. Information Systems **3**(6) (1981) 205–218
15. United States National Library of Medicine: Introduction to MeSH - 2011. `http://www.nlm.nih.gov/mesh/introduction.html` (2011)
16. Roberts, R.J.: PubMed Central: The GenBank of the published literature. Proceedings of the National Academy of Sciences of the United States of America **98**(2) (January 2001) 381–382