Automated Similarity Modeling for Real-World Applications

Rotem Stram

Knowledge Management Group, German Research Center for Artificial Intelligence, Kaiserslautern, Germany
rotem.stram@dfki.de

Abstract. Many Case-Based Reasoning (CBR) applications rely on experts opinions and input to design the knowledge base. Even though these experts are an integral part of the modeling process, they are human and cannot always provide the amounts of information that is needed. To that effect, the idea behind this thesis research is to utilize the data's struc-ture to extract relationships between knowledge entities in cases where expert knowledge is not enough. The goal is to automatically model the similarity measure between cases and their attributes using methods such as information retrieval (IR), natural language processing (NLP), machine learning, graph theory, and social network analysis (SNA), with an emphasis on SNA, to extract contextual knowledge from a dataset.

Keywords: Similarity, Social Network Analysis, Graph Theory, Sensitivity Analysis, Machine Learning, Information Retrieval

1 Introduction

In the heart of CBR is the similarity measure between two cases. Several successful models have been used to compare cases, both locally and globally. Most notably are taxonomies and matrices for local similarities of attribute-value-type fields, and weighted sum for global similarities of entire cases. What these methods have in common is that they rely on a pre-existing similarity value between two items or concepts. These are traditionally modelled by experts in the domain of the system.

Statistical methods have been used in the past, mainly in Textual CBR, to measure the proximity of two items. Methods such as TF-IDF and cosine similarities have been used successfully in the IR field, and augmented with additional information when used in CBR. These augmentations include abbreviations, synonyms, taxonomies, and ontologies as provided by experts [4].

Practical examples include SCOOBIE, where subject-predicate-object triplets in the form of RDF graphs are extracted from text, with the help of a pre-existing domain-specific ontology supplemented with linked open data such as DBpedia [7]. In KROSA a pre-existing ontology was used to extract requirements from

text. Here phrases and words were obtained using NLP methods and matched against items in the ontology [3]. These approaches have poor adaptability, since extending the vocabulary requires a lot of effort both in term acquisition and similarity measures.

Other examples allow better adaptability to changing and expanding vocabulary. One such approach is Probst et al., where an attempt was made to extract attribute-value pairs from text. Seeds were used to train a model to classify noun phrases in a semi-supervised manner. However, this work relies on texts with a predictable structure and does not describe how the extracted values relate to each other [5]. In Bach et al. terms are extracted from text with basic NLP techniques, and are then assigned to classes by experts, who also model their similarity [1].

When dealing with real-world data it is practically impossible to model all items that may occur, due to the usually large amount of data and the different identities of the contributors. Expert input is an important tool to identify the main concepts of a domain and have an understanding of how they relate to each other, but cannot possibly cover the entire extent of actual values of an attribute and the relationships between those values.

In the scope of the OMAHA project for fault diagnosis in the aircraft domain [9], we are dealing with a large semi-structured data set of faults and their solutions. Information about the aircraft and the faulty system is given in the form of symbolic attributes, while the fault description is in natural language form, as written by the technicians on site. We have been provided with a list of abbreviations, taxonomies of major concepts, and in the future white- and black-lists. WordNet is being employed for synonyms, but is too general for our purposes.

2 PhD Research Focus

It is the goal of the PhD thesis research to automatically identify important concepts in the corpus, and to model the relationships between them. A balance should be reached between expert knowledge and machine learning, taking both into account. Since an expert can assist in finding the main concepts of a corpus and the relationships between them, but cannot predict all possible values, both present and future, this process needs to be automated so that the system can evolve and develop without constant supervision. Methods on which the thesis may focus on are IR, NLP, graph theory, SNA, and machine learning.

The current area of research is the definition of similarity of symbolic attributes, with a potentially unlimited number of values. These values originate from textual representation of cases, and are extracted using IR and NLP methods. They are then ordered into different attributes of a case. With this in mind, cases and values can be seen as nodes in a social network, connected between them and with each other. This allows the utilization of SNA methods to model this interaction and extract information about it. Using SNA to model similarity

is rooted in the idea that nodes that share a similar environment become similar [2]. The knowledge gained can then be used as a starting point for machine learning methods to optimize global similarity.

The OMAHA project is reaching its final stages, but there is still much to be done within the scope of automatic acquisition of similarity. The measures should be applicable to a wide range of scenarios, and not specific to the given project. The research is still in a very early stage, and a focus has not been decided yet.

3 Current Progress

A few concrete steps have been taken to model similarity within the OMAHA project, and in general. The following will give a short description.

3.1 Global Similarity with Sensitivity Analysis

A new method of global similarity assessment has been developed based on sensitivity analysis of case attributes to the corresponding diagnosis, and the paper on the matter has been accepted for publication [8]. The idea is that in order to distinguish between different diagnoses, different attributes may be more important. For instance, in a set of diagnoses, deciding if a case belongs to a specific one it would be most beneficial to look into the value of attribute a. If it does not belong to it, then attribute b may play a key role in deciding whether it belongs to a different diagnosis, and so on.

The sensitivity analysis combines a statistical analysis of attribute values and their association with each diagnosis, together with a learning stage where weights are learned by supervised learning. During this stage a set of retrievals are performed and their outcome analyzed. The weights are then updated according to their contribution to the retrieval of irrelevant cases. In the end a weight matrix is outputted, giving each attribute a weight under each diagnosis. This method builds on Richter and Wess' work, and expands it to include any type of attribute [6].

3.2 Natural Language Processing

This is a means to an end, mainly to obtain values for the attribute. Since the fault description text we were given was written by experts in the field, who were probably in a hurry and whose English is not their native language, POS tagging yielded no coherent results. It was needed instead to manually create patterns with the help of regular expressions and phrase extraction. This part is dataset-specific.

3.3 Similarity Assessment with SNA

After concepts were extracted with the help of NLP, the dataset was regarded as a bipartite graph, connecting concepts with diagnoses. Using one-mode projection, a weight between each attribute value was calculated and regarded as the similarity value. This method is still being tested and will be extended in the future.

Since we are dealing with potentially unlimited (but not infinite) number of concepts per attribute, it is very likely that application domains have a long tail of concepts that are used in very few cases. SNA allows to include those concepts as part of the whole system, creating an asymmetrical similarity measure that describes the connection of each value to another.

References

1. Bach, K., Althoff, K. D., Newo, R., Stahl, A. A case-based reasoning approach for providing machine diagnosis from service reports. In International Conference on Case-Based Reasoning (pp. 363-377). Springer Berlin Heidelberg. 2011.
2. Borgatti, S. P., Mehra, A., Brass, D. J., Labianca, G. Network analysis in the social sciences. science, 323(5916), 892-895. 2009.
3. Daramola, O., Stlhane, T., Omoronyia, I., Sindre, G. Using ontologies and machine learning for hazard identification and safety analysis. In Managing requirements knowledge (pp. 117-141). Springer Berlin Heidelberg. 2013.
4. Lenz M.: Textual CBR and Information Retrieval A Comparison. Proceedings 6th German workshop on CBR. 1998.
5. Probst, K., Ghani, R., Krema, M., Fano, A. E., Liu, Y. Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions. In IJCAI (Vol. 7, pp. 2838-2843). 2007.
6. Richter, M. M., Wess, S. Similarity, uncertainty and case-based reasoning in PATDEX. In Automated Reasoning pp. 249-265. Springer Netherlands. 1991.
7. Roth-Berghofer, T., Adrian, B., Dengel, A. Case acquisition from text: Ontology-based information extraction with SCOOBIE for myCBR. In International Conference on Case-Based Reasoning (pp. 451-464). Springer Berlin Heidelberg. 2010
8. Stram R., Reuss P., Althoff KD., Henkel W., Fischer D.: Relevance Matrix Generation using Sensitivity Analysis in a Case-Based Reasoning Environment. ICCBR (accepted). 2016.
9. German Aerospace Center - DLR, LuFo-Projekt OMAHA gestartet, http://www.dlr.de/lk/desktopdefault.aspx/tabid-4472/15942_read-45359/ (last followed on July 25, 2016)