Poster presented at the TRIO/ITRC Research Retreat, Queen's University, May 10-12, 1994

Similarity Assessment

Igor Jurišica[1]
juris@cs.utoronto.ca
Department of Computer Science
University of Toronto
Toronto, Ontario M5S 1A4, Canada

In many situations, one cannot locate an exact information. It may be either because such information is not available or because approximate information is required in addition to the exact one. The obvious problem is how to find these approximations and how to measure them. One way to locate an approximate information is to define a similarity measure and to use similarity-based retrieval. Even though several approaches had been developed in the past, there still are problems to be solved.

One problem of similarity-based retrieval is based on the fact that what is similar in one situation may not be similar in another. Thus, the similarity measure should not be predefined in the system. Instead, it should be specified at query time. This would allow for flexibility by specifying what parts of the knowledge are relevant for assessing similarity in the current situation.

Another problem is judging the relevance of the retrieved information to the current needs. Most of the systems return the information which is the closest approximation of the requested one (usually measured by a number of matching features). However, this says little about the usefulness of the information. Similarity assessment should support a mechanism for judging the usefulness of retrieved information.

Our approach to flexible similarity assessment assumes a new definition of similarity. The key part of this definition is the presence of context. Context in similarity allows for: retrieving a "hard-to-locate" information, a flexible similarity assessment, judging the relevance of the retrieved information, and finding circumstances where symmetry and transitivity in similarity hold.

We present an introduction to the new theory of similarity with focus on the formal definition of context and similarity. Similarity with context is defined in terms of set theory and is a basis of a theoretical foundation for flexible similarity-based retrieval. Similarity is a function with three parameters: a set of objects (SO), a context (Ω), and an information base ($\Delta \supseteq SO$). Based on the presented formal notation, we present four variations of the similarity applicable in different tasks, namely: similarity in comparison, similarity in retrieval, similarity in evaluation, and similarity in analysis.

For *similarity in comparison*, both context and the set of objects are specified. *Similarity in retrieval* is used when the context is specified and the task is to retrieve all relevant objects from the information base; thus, the set of objects is a variable. A possible use of similarity in retrieval is when one has to locate all relevant objects in an information base and it is impossible to locate them directly. In such a situation, it is necessary to specify a certain context which should be used to locate relevant (similar) objects. Context specifies important features used to describe an object. The purpose of *similarity in evaluation* is to evaluate the similarity between given objects by finding possible contexts; hence, context is a variable. In the real world, however, the objective is to find the most restrictive context (since there is an infinite number of other possibilities). *Similarity in analysis* can be considered as a way of knowledge mining, i.e., a way of finding interesting relations within an information base.

[1]Supported by the University of Toronto Open Doctoral Fellowship and a Research Assistantship from the Department of Computer Science