Development of an Information Retrieval Method Based on the Rough Set Theory

Kaname FUNAKOSHI

School of Information Science,
Japan Advanced Institute of Science and Technology

February 14, 1997

Keywords: information retrieval system, rough set theory, tolerance relation.

Information Retrieval

Information retrieval is a branch of information processing which aims at searching documents from the database suitably for the user's interest. There are numerous bibliographical data of documents stored in huge databases and the number of documents is growing day after day. An ideal information retrieval system which provides the list of appropriate documents to the user is desirable but it is very difficult to construct such a system.

An information retrieval system can be formulated as $\mathcal{S} = (\mathcal{T}, \mathcal{D}, \mathcal{Q}, \alpha)$ where $\mathcal{T}$ is a set of keywords, $\mathcal{D}$ is a set of documents where each document is a subset of $\mathcal{T}$, $\mathcal{Q}$ is a set of query where each query is a subset of $\mathcal{T}$ and $\alpha : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}^+$ is a retrieval function which indicates the relevance between the document and the query. When the user gives a query Q, the system $\mathcal{S}$ will retrieve relevant documents in $\mathcal{D}$ to Q with the corresponding value of α.

Almost information retrieval systems use Boolean operations for conceptual models because of their simplicity, but Boolean operations cannot always provide suitable answer to the user interest. There have been attemps to improve information retrieval quality by non-Boolean models. Intelligent matching strategies for information retrieval often use the concept analysis requiring semantic calculations. The rough set approach to information retrieval investigated in this work is also in this direction.

Rough Set Theory and Tolerance Relations

The rough set theory, introduced by Pawlak in early 1980s, is considered as a good mathematical tool to deal with vagueness and uncertainty. The primary notions of this theory are *approximation space* and *lower* and *upper approximations*. Consider the universe U as a set of objects described by a set of attributes. U can be partitioned into a set of *equivalence classes* by using an *equivalence relation*. Two elements $x, y \in U$ belonging to the same equivalence class are considered indiscernible. This partitioned universe is called *approximation space*. For any subset $X \subseteq U$, the *lower approximation* of X, denoted by $\mathcal{L}(X)$, is the union of approximation classes which are included in X and its *upper approximation* denoted by $\mathcal{U}(X)$ is the union of equivalence classes which have non-empty intersection with X.

The equivalence relation requires the reflexive, symmetric and transitive properties. However, the transitive property does not hold in certain application domains such as natural language processing and information retrieval as semantics of terms in documents are not transitive. Information retrieval systems based on the rough set theory were investigated by Srinivasan late in 1980s in which they used rough equivalence relations.

A generalized rough set theory by using tolerance relations which does not require the transitive property was introduced in 1994 by Skowron and Stephaniuk. The tolerance relation classifies elements in the universe into groups called *tolerance classes*. It is known that equivalence classes are disjoint but tolerance classes can be overlapped. The universe grouped by tolerance classes is called *tolerance space* and denoted by $\mathcal{R} = (U, I, \nu, P)$ where U is the universe, $I : U \rightarrow \mathcal{P}(U)$ is an uncertainty function, $\nu : \mathcal{P}(U) \times \mathcal{P}(U) \rightarrow [0, 1]$ is a vague inclusion function, and $P : I(U) \rightarrow \{0, 1\}$ is a structurality function. Each element $x \in U$ has a tolerance class denoted by $I(x)$. For any subset $X \subseteq U$, its lower and upper approximations are defined as $\mathcal{L}(\mathcal{R}, X) = \{x \in U \mid \nu(I(x), X) = 1\}$ and $\mathcal{U}(\mathcal{R}, X) = \{x \in U \mid \nu(I(x), X) > 0\}$. By using this generalized rough set theory, we have developed an information retrieval model.

An Information Retrieval Model Using Rough Tolerance Relation

In this study we propose a conceptual model for information retrieval based on the rough set theory by using tolerance relations instead of equivalence relations. The work consists of three parts: determination of the tolerance space, definition of rough inclusions (including equality and overlap) and retrieval algorithm, the system implementation and evaluation.

For any subset $X, Y \subseteq U$, three relations are defined. If the lower/upper approximations of X and Y are the same, they are called *roughly equal*; if the lower/upper approximations of a set is included in the lower/upper approximations of the other, they are called *roughly included*; and if the lower/upper approximations of X and Y are overlapping, they are called *roughly overlap*. There are three kinds of rough equality and rough inclusion corre-

sponding to the cases when only lower approximations are roughly equal/included, only upper approximations are roughly equal/included, and both lower and upper approximations are roughly equal/included. There are also two kinds of rough overlap corresponding to the cases when only upper approximations are overlapping and both lower and upper approximations are overlapping. Totally, there are 8 kinds of rough tolerance inclusions.

In this work we use the whole set of keywords in database as the universe. While documents d_j and query Q are both sets of keywords and then subsets of the universe, the documents are selected by the rough tolerance relationships between d_j and Q. Actually, the matching is done in five layers: (1) exact match, (2) rough equalities, (3) rough inclusions (query is roughly included in document), (4) rough inclusions (document is roughly included in query), and (5) rough overlaps. Each document is checked from layer (1) to (5) to see they satisfy which layer. Finally, five layers of documents are retrieved. These layers are divided detailly into 12 levels of inclusions. The process is formulated in an algorithm.

The co-occurrencies of keywords in the database are used to determine the tolerance relation in our model. Denote by $c(t_i, t_j)$ the total number of co-occurrencies of two terms t_i and t_j. If $c(t_i, t_j) \geq \theta$ then t_i and t_j are regarded as in the same tolerance class where θ is a threshold.

The secondary ranking method is introduced to the layer (5) for improving the retrieval effectiveness as it may prevent the case that too many documents can be retrieved in a single layer. This new ranking method consists of a simple intersection calculation between the document and the query.

Implementation and Evaluation

The method is implemented and evaluated through a real database. The system consists of two phases: calculating the tolerance classes of all keywords, the lower and upper approximations of all documents in the database; and searching documents relevant to a given query. All documents of the Journal of Japanese Society for Artificial Intelligence in first 10 years (1986-1995) are used as the database. Keywords determined by the authors are used for characterizing documents. There are 802 documents and 1813 keywords in the database.

The system can provide well ordered documents in the layers (1), (2), (3) and (4) with very high precision. The layer (5) often contains a large number of documents. By the secondary ranking method, the precision of the method is evidently increased though the recall is not much decreased.

The main advantage of the rough tolerance model for information retrieval is the ability of searching documents with high precision. The rough overlap searching is sometimes not convenient to a proper retrieval. The model presents encouraging features in information retrieval and worth to be further investigated.