Abstract

Classification is an integral part of an information retrieval environment. Clustering techniques are employed in information retrieval systems to enhance the effectiveness of the retrieval process. In this paper, we advance a knowledge-based approach to classification. The proposed methodology employs Personal Construct Theory for interviewing a domain expert to elicit classification knowledge. The knowledge elicited from the expert is mapped to system observable features of documents to develop a classification. The techniques developed are experimentally validated.

1 Introduction

Clustering techniques are employed in an information retrieval (IR) system to improve the efficiency and effectiveness of the retrieval process [5]. Clustering comprises of a mechanism to systematically arrange documents such that related documents are placed close to each other within the storage medium. In addition, a clustering scheme should provide capability to assign new documents to the most appropriate cluster.

In earlier IR systems, clustering was based on an association measure between document descriptions computed as a result of an indexing process [5]. However, it was difficult to accomplish an ideal classification that preserved users' viewpoint of association between documents. Users of an IR system may view different terms in different perspectives. This affects the recall as some relevant documents that do not conform to the user vocabulary may not be retrieved. Therefore, it was found desirable that the users be able to formulate a query in their own vocabulary. These observations led to the development of adaptive IR systems [6].

The adaptive systems observe user behavior with respect to different queries over a period of time and use this observation as a basis for the construction of clusters within the collection. However, addition of new documents necessitates the reclustering of the entire collection instead of the document being assigned to an appropriate cluster. This is because the system needs to observe the user behavior with respect to the new documents vis-a-vis the existing collection.

In the proposed model, the classification is achieved by formulating rules that capture the user biases. A proper quantification of uncertainty associated with each rule facilitates the ranking of appropriateness of clusters with respect to a new document. This quantification of uncertainty is a hard problem to solve. Generally, a user has to experiment by assigning different values to the uncertainty in the rules before an optimal rule-base can be formulated [2]. An automated knowledge acquisition tool employing structured interview techniques can be used to quantify the uncertainties and construct the rule-base for classification.

In this paper, we investigate the use of Personal Construct Psychology to automatically interview a user. This interview results in raw data which, on analysis, yields the relationship between different concepts from user perspective. After finding the relationships, the user is asked to delineate the boundaries which enclose like concepts. With such a grouping of concepts, we develop a methodology to establish a relationship between the concepts and the index terms constituting document representations. This relationship is employed to assign a document to the most appropriate cluster.

*This research is supported by the NSF grant IRI-8805875.
2 Personal Construct Theory

Personal Construct Theory is employed by psychologists to assist their clients in classifying different objects and persons (entities) encountered by the clients in their everyday life [4]. The properties of these entities make the person behave in a characteristic manner and are known as constructs. A psychologist needs to determine the combination of constructs and entities which lead to the peculiar behavior of the person.

The pertinent information is determined from the client through an interview. Before the interview, the interviewee enumerates the entities. Three of those entities, selected at random, are presented to the interviewee who then names some construct that sets two members of this triad apart from the third. The interviewee then assigns a rating, on a predetermined scale, to all the entities according to the confidence with which he perceives the construct to be present in the entity. The process is repeated many times by random selection of a new triad of entities. During the interview, the interviewee can also volunteer impromptu some constructs or additional entities.

The rating of all entities on each construct forms a matrix known as repertory grid. In repertory grid, each row and column correspond, respectively, to a construct and an entity given by the interviewee. The item in the ith row and jth column gives the rating of jth entity over ith construct. The repertory grid contains a person's viewpoint in the form of raw data and provides a ready cross reference between entities and constructs. An analysis of this raw data brings forth the relationships between different entities or constructs while retaining the viewpoint of the interviewee. These relationships form the basis for classification.

3 Concept Elicitation in Information Retrieval

By the very nature of personal construct theory, the constructs elicited from a person are unique to him. This can be exploited to determine the concepts deemed important by a person to satisfy his information needs. A set of documents well known to the user of an IR system are selected and treated as entities. This set is called training set. The user identifies the constructs (constructs) which he considers important in the description of documents in the training set. The documents in the training set are rated on each concept by the user on a predetermined rating scale. The ratings are collected in the repertory grid which is analyzed to explicitly determine user biases.

Let the constructs elicited from the user be denoted by c1, . . . , cn, and the documents in the learning set be identified by d1, . . . , dm. If the rating scale varies between 1 and k, the extent of mutual relationship between two constructs ci and cj is quantified by the expected mutual information measure (EMIM) I given by

$$I(c_i, c_j) = \sum_{u=1}^{k} \frac{\sum_{v=1}^{k} P(c_i = u, c_j = v)}{\log \frac{P(c_i = u)P(c_j = v)}{P(c_i = u, c_j = v)}}$$

(1)

where $P(c_i = u)$ is the apriori probability associated with the event that a construct has the value u assigned with respect to the ith construct and $P(c_i = u, c_j = v)$ is the conditional probability of document $d_i$ being rated u on the construct $c_i$ when $d_i$ is rated v on construct $c_j$ [3].

Since EMIM is symmetric, the EMIM between every pair of constructs can be tabulated in the form of an upper triangular matrix without any loss of information. This matrix is known as similarity matrix [1]. The diagonal elements in the similarity matrix describe the relationship of a construct to itself and are ignored.

The similarity matrix contains the relationship of each construct with every other construct. Therefore, in graph-theoretic terms, it corresponds to a complete graph which has the constructs as its nodes and the degree of dependence between different constructs as the weight of the edges between their corresponding nodes.

The complete graph can be reduced to a simpler graph which contains all the nodes of complete graph but retains only the most significant edges (relationships). Spanning trees provide a convenient representation for such a graph. A spanning tree is identified from the complete graph such that the total weight of arcs included in the spanning tree is optimal in a well-defined sense. This tree is known as a maximum weight dependence tree [3]. Let T be the set of all possible spanning (dependence) trees for a particular complete graph. A maximum weight dependence tree is a dependence tree $T \in T$ such that the total EMIM for $T$ is maximized with respect to all dependence trees $T' \in T$.

$$\sum_{i=1}^{n-1} I_T(c_i, c_{j(i)}) = \max_{T' \in T} \left( \sum_{i=1}^{n-1} I_{T'}(c_i, c_{j(i)}) \right)$$

(2)

where $j(i)$ is a function mapping a node $t_i$ into its adjacent node in the dependence tree under consideration and n is the number of constructs (nodes) in the tree.

The maximum weight dependence tree is presented to the user who is then asked to identify groups of con-
structs such that each group has some traits common to all the constructs in the group. In each group, the user identifies one construct that most closely identifies the group as a whole. This construct is known as the super-ordinate construct $C$ for the group. Each super-ordinate construct embodies a group of constructs and is given by

$$C_i = \{c_{i1}, c_{i2}, \ldots \} \quad (3)$$

Also, each construct $c_i$ identified by the user during repertory grid elicitation interview belongs to only one class such that the classes do not overlap.

$$\{c_i \in C_i \land c_j \in C_j, \land C_i \neq C_j\} \Rightarrow c_i \neq c_j \quad (4)$$

4 Development of Classification Rules

This section describes the development of classification rules employing the concepts elicited from the user. The concepts are aggregated into classes and rules formulated to assign the documents to different classes.

Every object has a set of attributes associated with it such that the object is represented by this set. The attributes in this representation are known as the primitive attributes of the object with respect to the domain of application. Within the domain of IR, these primitive attributes are keywords that describe the documents.

To classify new documents, an appropriate representation needs to be developed for each concept as well as the superordinate construct. The representation for concepts is developed by determining the discriminating terms for the concepts from the representation of documents.

Each concept identified by the domain expert induces a partition on the learning set. As a result of this partition, a document in the learning set belongs to one or the other equivalence class. For a given concept, two documents in the learning set are said to belong to the same equivalence class if and only if they are assigned the same rating for that concept. The equivalence class induced by a concept $c_i$ with rating $k$ is represented by

$$[c_i]_k = \{d_t \mid r_{it} = k\} \quad (5)$$

A representation for each document is developed by employing the automatic indexing techniques. This process results into a set of index terms or keywords that adequately describe the contents of the documents. Thus, document $d_t$ is represented by

$$d_t = \{t_{t1}, t_{t2}, \ldots \} \quad (6)$$

The representation for an equivalence class is developed by selecting the discriminating terms for the documents in that class. These terms are meant to be the ones that are present in the documents belonging in the class but not present in the documents that are not in the equivalence class.

$$[c_i]_k = \text{d}_t - \text{d}_t' \mid r_{it} = k, r_{it'} \neq k \quad (7)$$

In other words, if $d_t \in [c_i]_k$ and $d_t' \notin [c_i]_k$, then,

$$[c_i]_k = \text{d}_t - \text{d}_t' \mid t_i \in d_t, t_i' \notin d_t' \quad (8)$$

Now, it is easy to develop the representation for a group of concepts, denoted by the super-ordinate concept $C$.

$$[C]_k = \text{d}_t - \text{d}_t' \mid r_{it} = k \text{ for some } c_i \in C, \quad r_{it'} \neq k \text{ for any } c_i \in C \quad (9)$$

The actual rules for classification are developed as follows. Let $\circ$ denote a suitable matching function and $d$ be the document to be classified. A representation for the document is developed employing classical indexing techniques such that

$$d = \{t_1, t_2, \ldots \} \quad (10)$$

The document is matched against the representation of each group of concepts and is assigned to the group for which the matching function is maximized.

$$d \in C_i \Leftrightarrow d \circ C_i = \max_{\forall C_j} (d \circ C_i) \quad (10)$$

5 Experimental Results

To demonstrate the effectiveness of our approach, an elicitation experiment was performed with a training set of 24 documents, broadly consisted of documents on the topics of “knowledge acquisition,” “intelligent tutoring systems,” “information retrieval,” and “expert systems.” Each document in the collection was represented by its complete bibliographical reference and abstract. The repertory grid and the similarity matrix from this experiment are described in [1]. The maximal spanning tree identified from the similarity matrix is displayed in Figure 1. The shaded areas in the tree represent the classes of constructs identified by the expert.

Using the keyword representation for documents in the learning set and expression 9, the representation for each class was developed employing a Prolog program.
To validate the techniques, a collection of 122 documents was used. The training set was previously selected from this collection. The keyword representation for each document in the collection was developed using conventional indexing techniques. The documents were classified by a Prolog program employing the rules developed and the cosine matching function [5]. The approach used for validation is to check the efficacy of the procedure in actual classification.

The representation of each document is matched against the representation of groups of constructs. The document is assigned to the group for which this matching is maximized. The result of this assignment is displayed in Table 1. Table 1 shows that more than 80% of the documents are correctly classified. 4% of the documents are not classified because of inadequate number of classes. This is expected because these documents are not typical of any class in the collection and were not represented by any document in the learning set. About 14% of the documents were classified into incorrect classes. However, the documents has only a small amount of difference in the value of matching coefficients due to correct and assigned class.

6 Discussion and Future Work

In this paper, we have advanced the idea of using knowledge acquisition techniques to develop a classification environment. The main advantage of this technique is a quick transfer to the new system with a minimal disruption and a minimal set up time while preserving the existing classification. The techniques developed were experimentally validated by implementing a classification system. The experimental results are encouraging. A comparative evaluation of the knowledge-based classification scheme and existing classification techniques is in progress. Improvement to the development of cluster representation by employing learning algorithms is also being investigated. It is expected that the application of learning algorithms will enhance the overall classification environment and the effectiveness of classification.

References


