AIDOS Software AG

KAI| Box Technology

The Categorization System

Inhalt

1 Taxonomy and Categorization

KAI| Box prepares all registered information so that the search module optimally supports the user during the search for the required information.

The automatic categorization is a module which prepares the information. It results in the classification of the registered documents in a company-specific taxonomy which is constructed by the knowledge administrator and trained on the basis of a set of training documents. User-specific category models will also be supported in the near future.

Taxonomy stands for a hierarchical structure of categories and subcategories in which objects are arranged according to certain characteristics. It can illustrate the structure of the company, contain certain areas of responsibility, include the competition, etc. It is essential for the taxonomy that documents which are assigned to a subcategory also belong to the respective superordinated category.

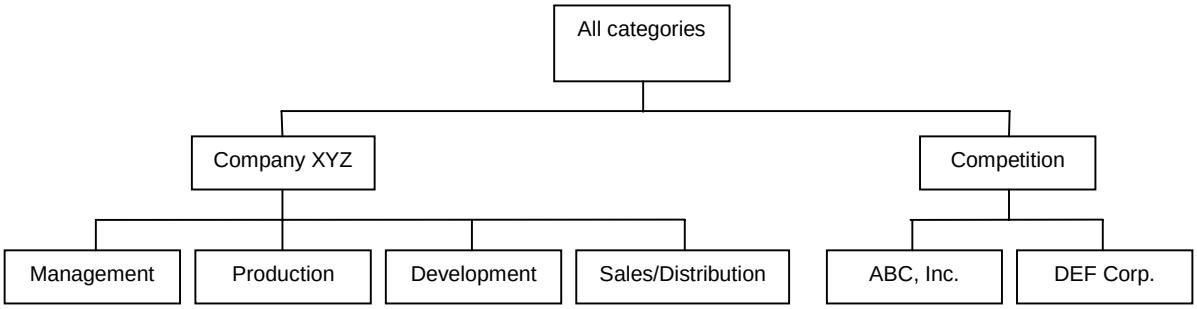

Fig. 1-1 Example of a Taxonomy

For the user, the advantage of automatic categorization mainly lies in the fact that the search results can be displayed indexed by categories. This enables the user to choose to only see those documents which are assigned to the categories most interesting for him. By selecting his catetories of interest at the beginning of the search process, the user can restrict the amount of documents to be searched through.

The graphical illustration using the visualization module assists the user with orientation within the available taxonomy.

2 The Categorization Module

2.1 Description of the Automatic Categorization

The categorization of documents using the categorization module is based on a progressive process known as „categorization by example“.

The system „learns“ to distinguish between the single classes of a taxonomy by means of a set of training texts. A category model to categorize unknown texts is developed from the unity of the taxonomy and the training set. Changes in the training set entail changes of the category model itself. The result of a categorization of the same text before and after a change of the training set – each time requiring new training and tuning – can be different. This is usually intended to increase the quality of the categorization.

The tuning of the category model is used to set intrasystem values which optimize the exactness of the categorization process. For that purpose, the precision and recall values – further details in section 2.2 – are determined for each category for the „high confidence“ and low confidence“ cases, since both values affect each other reciprocally.

After the training and tuning process, the category model is then stored. The categorization of new incoming documents is now based on this model using the kNN method described in chapter 3. The resulting output consists of the determined categories, also indicating the degree of confidence and the "implData" value. Value "1" stands for "high confidence", "0" for "low confidence". The "implData" value is a non-standardized value. The higher it is, the more similar the categorized document is to the determined class. This value however, is not comparable for different categories.

2.2 The Precision and Recall Values

In the field of information processing, the precision and recall values were developed as a measure for the evaluation of automatic systems for information retrieval. They represent the correctness and completeness of such a system.

Under the assumption that an information retrieval system delivers a variable number of replies to one inquiry, some of which are wrong and some of which are correct, these terms are defined as follows:

Precision:	Relation of the number of correct replies to the total number of replies
Recall:	Relation of the number of correct replies to the number of all possible correct replies.

There are reciprocal effects between precision and recall. A high value for precision entails that only those results are determined, which are very probably correct, so that fewer replies are given. This results in a low recall value. A low precision value on the other hand increases the recall value. These values depend on the degree of correspondence.

Example:

The process of buying a house can be divided into two phases. In the first phase, one looks at houses and wants to see the largest possible selection of all available houses. This means that the recall value is more important. In the second phase, when the decision as to which house to buy takes place, the precision value is more significant. After all, one wants to buy the right house. In this case, the confidence value describes the correspondence of the characteristics of all available houses with the characteristics of the chosen house. The following diagram illustrates the connection between precision and recall depending on confidence:

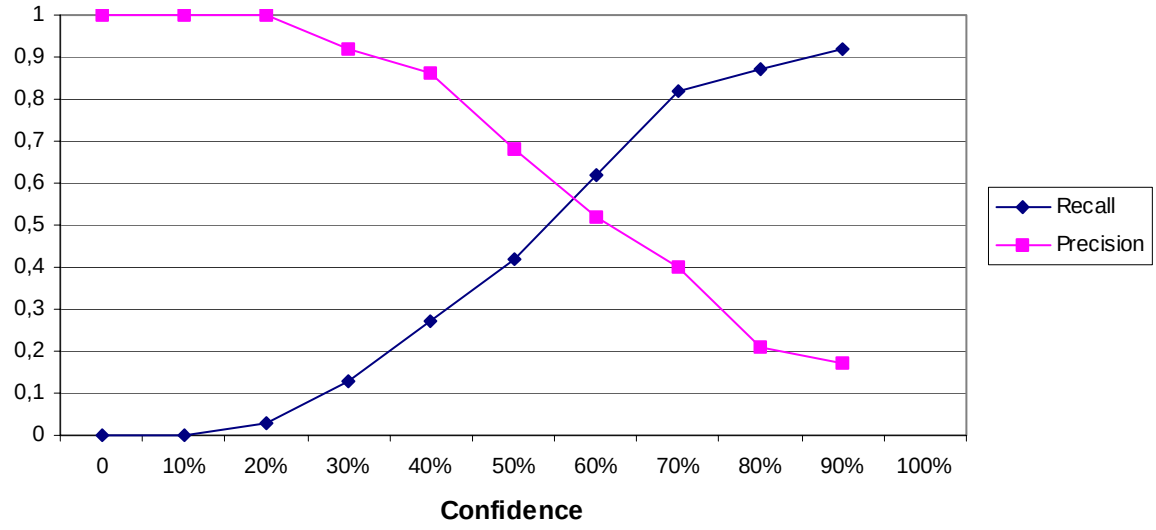

Fig. 2-1 Connection between precision and recall

In order to find an optimum between precision and recall, in the area of information retrieval two main values are used: the f-score and the beta-value. F-score measures the total precision of the system as a combination of precision p and recall r. The calculation takes place according to the formula:

$$f = ((ß^2 + 1)\, p\, r) / (ß^2\, p + r)$$

The optimization of the system influences the balance between precision and recall by varying the ß value. For ß = 1, precision and recall are evenly balanced, for ß < 1, precision is emphasized, for ß > 1, this is done for recall.

This process takes place automatically and without the user's influence.

3 The kNN-Method

The KNN-method – the „k nearest neighbors“ function – is a statistical method. It is used to determine the affiliation of an object to a class from a number of several possible classes. The possible classes are determined by the objects assigned to them. For the unknown object, k objects (this stands for a fixed number) with the smallest distance to the unknown object are determined. The object to be classified is assigned to the class which contains most of the „k nearest neighbors“.

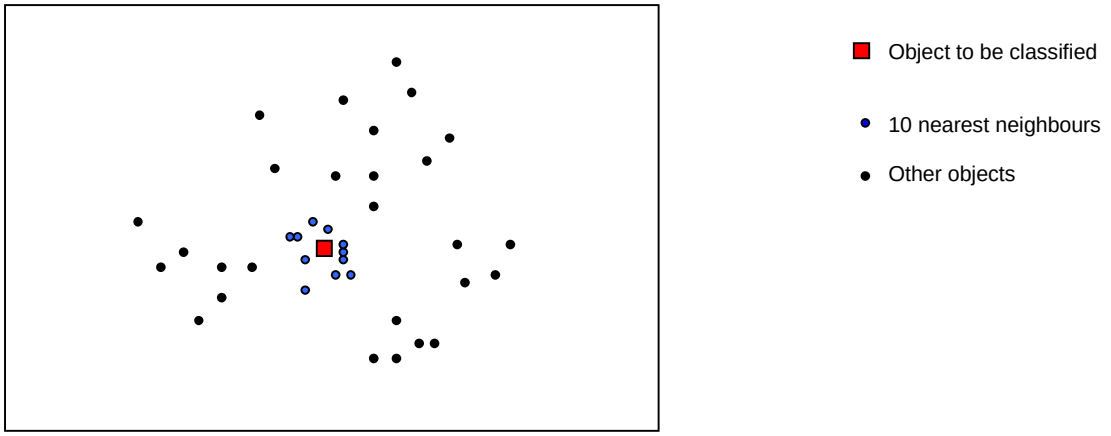

Fig. 3-1 The 10 nearest neighbours

During the training, the documents of the learning set are mapped on a n-dimensional vector space. The mapping is done based on all the words contained in the documents with reference to the total amount of words in all training documents. A linguistic analysis of the texts precedes this process. The vector space contains a defined distance function. This distance function is used to calculate the distance from one vector to the others.

During its categorization, a document is also mapped on the vector space, and the distance to the other documents of the learning set is determined. As described above, in general, this allows us to determine the „k nearest neighbors“ of the document set. The „k nearest neighbors“ therefore are the „k most similar documents“ of the learning set. The natural number k is defined to be 60 as a standard. The size of the number k has further implications on the expected quality of the categorization in connection with the number of training documents.

Note:

A document is only determined as a nearest neighbor to the document to be categorized if the document to be categorized also belongs to the k nearest neighbors of the other document. This relationship has to be valid in both directions.

As a next step, the categories are determined for each of the identified nearest neighbors which the document in the learning set was assigned to. The document is then assigned to the category which most of its nearest neighbors were assigned to as well. Since during this process in the KAI| Box categorization module, a document can be assigned to several categories, the selection of the appropriate category is based on intrasystem rules (threshold values).

The following three steps briefly describe the process:

1. The categorizer finds the k most similar documents to the document to be categorized from the learning set. (standard value 60)

2. The categorizer determines the category assigned to each of the k most similar documents.

3. The categorizer examines every determined category and assigns the document to it if this category is assigned to a a sufficient amount of similar documents. At the same time, the categorizer determines a „confidence“-indicator ("implData"-value) which is based on the number of the most similar documents which belong to this category.

4 Administration of the category model

The *KAI| Box* category model administrator enables the knowledge administrator to develop and elaborate upon a company-specific category model as well as adjust it to developmental changes. He can carry out changes in the taxonomy – remove or add categories/subcategories or modify hierarchical relations. The training sets can be modified the same way: new documents are added, while documents which are not typical anymore are deleted.

These changes to the category model can be carried out independently from the category model implemented in KAI| Box. Before a new or modified version of the category model is implemented, it can be sufficiently tested and optimized.

5 Table of Figures