Enhancing Short Text Retrieval in Databases

N. Marín, M.J. Martín-Bautista, M. Prados, and M.A. Vila

Intelligent Databases and Information Systems Research Group
http://idbis.ugr.es
Dept. of Computer Science and Artificial Intelligence
University of Granada, 18071 - Granada (Spain)
{nicm, mbautis, prados, vila}@decsai.ugr.es

Abstract. In this paper, we present a mechanism to deal with short text structures in relational databases. Text fields are transformed into a special knowledge representation named AP-structure based on the Apriori algorithm of the mining area. Once the abstract data type is obtained, the text fields can be summarized, mined, and queried in a easy way. The operations to query these fields are the main aim of this paper.

Keywords: Semantic querying, short texts, AP-sets, frequent itemsets, knowledge structure.

1 Introduction

Textual fields are not easy to treat in databases when a discovery process is carried out. The only operations over these fields are the classical ones in databases for text attributes such as to ask for the content of the field, or search for the fields containing a certain word or pattern. However, the lack of a reference domain for these fields make difficult operations such as semantic querying, mining, data warehousing, etc..

One possible solution to this problem is to give a structure to text fields. Although traditional Information Retrieval allow us to index these fields, a semantical structure should underly them. For this purpose, we propose an intermediate structure called AP-set [6] based on the frequent itemsets obtained by the Apriori algorithm of the mining area [2]. These sets are obtained as follows: all combinations of items in a transactional database are generated and only those having a support above a threshold called minsupport are considered and called frequent itemsets [1].

From the AP-set, abstract data types (ADT) can be established to facilitate its management. With this knowledge representation, different querying operations can be defined.

In the following section, we define the AP-set and the AP-structure concepts, as well as some operations to manage them regarding querying issues. An experimental example with a medical database is shown in Section 3. The problem of dealing with this new structure, in the context of an Object Relational Database

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 613–624, 2006.

System (ORDBMS) is discussed in Section 4. Section 5 offers some query models to the new database, as well as some examples of these for our medical database. Finally, some conclusions and future work can be found in Section 6.

2 Definitions of the Knowledge Representation Structures

In this section we will formalize the ideas presented above by defining the mathematical structures which will be the basis for the formal representation of data. Firstly, we will establish the definition and properties of the sets of subsets which have the Apriori property [2], that we have called AP-Sets. Next, we will give the formal definition and properties of the underlying structure in the texts which is that of a set of AP-Sets.

In this paper we are concerned with querying problems, therefore we have only included here those theoretical questions which refer to this point. Other operations between sets and AP-Structures and between AP-Structures, as well as its properties, have been established in [6].

2.1 AP-Set Definition and Properties

Definition 1. AP-Set
Let $X = \{x_1 ... x_n\}$ *be any referential and* $\mathcal{R} \subseteq \mathcal{P}(X)$ *(parts of the referential). We will say that* $\mathcal{R}$ *is an AP-Set if, and only if:*

1. $\forall Z \in \mathcal{R} \Rightarrow \mathcal{P}(Z) \subseteq \mathcal{R}$
2. $\exists Y \in \mathcal{R}$ *such that* :
   (a) $card(Y) = max_{Z \in \mathcal{R}}(card(Z))$ *and* $\neg \exists Y' \in \mathcal{R} | card(Y') = card(Y)$
   (b) $\forall Z \in \mathcal{R}; Z \subseteq Y$

In the case of a text application, the referential would be terms and their frequencies.

The set Y of maximal cardinal characterizes the AP-Set and it will be called *spanning set of* $\mathcal{R}$. We will denote $\mathcal{R} = g(Y)$, that is $g(Y)$ will be the AP-Set with spanning set Y.

We will call *Level of* $g(Y)$ to the cardinal of Y. Obviously, AP-Sets of level equal to 1 are the elements of X and we will consider the empty set $\emptyset$ as the AP-Set of zero level. It should be remarked that the definition 1 implies that any AP-Set $g(Y)$ is in fact the reticulum of $\mathcal{P}(Y)$.

Definition 2. AP-Set Inclusion
Let $\mathcal{R} = g(R)$ *and* $\mathcal{S} = g(S)$ *be two AP-Sets with the same referential:*

$$\mathcal{R} \subseteq \mathcal{S} \Leftrightarrow R \subseteq S$$

Definition 3. Induced sub-AP-Set
Let $\mathcal{R} = g(R)$ *and* $Y \subseteq X$ *be. We will say* $\mathcal{S}$ *is the* sub-AP-Set induced by Y *iff:*

$$\mathcal{S} = g(R \bigcap Y)$$