Chapter 15
Search in Documents Based on Topical Development

Jan Martinovič, Václav Snášel, Jiří Dvorský, and Pavla Dráždílová

Abstract. An important service for systems providing access to information is the organization of returned search results. Vector model search results may be represented by a sphere in an n-dimensional space. A query represents the center of this sphere whose size is determined by its radius or by the amount of documents it contains. The goal of searching is to have all documents relevant to a query present within this sphere. It is known that not all relevant documents are present in this sphere and that is why various methods for improving search results, which can be implemented on the basis of expanding the original question, have been developed. Our goal is to utilize knowledge of document similarity contained in textual databases to obtain a larger amount of relevant documents while minimizing those cancelled due to their irrelevance. In the article we will define the concept $k$-path (topical development). For the individual development of vector query results, we will propose the SORT-EACH algorithm, which uses the aforementioned methods for acquiring topical development.

Keywords: Topical Development, Clustering, Information Retrieval.

15.1 Introduction

There are many systems used for searching collections of textual documents. These systems are based on the vector model, probability models and other models for document representation, queries, rules and procedures [2, 19]. All of these systems contain a number of limitations. Incomplete lists of relevant documents obtained in search results ranks among one of the most basic of these limitations.
An important service for systems providing access to information is the organization of returned search results. Conventional IR systems evaluate obtained documents based on their similarity to given query. Other systems present graphic illustrations based on mutually similar documents, specific attribute relations and samples of terms distributed in the query.

Vector model search results may be represented by a sphere in an $n$-dimensional space. A query represents the center of this sphere whose size is determined by its radius (range query) or by the amount of documents it contains (NN–query). The goal of searching is to have all documents relevant to a query present within this sphere. It is known that not all relevant documents are present in this sphere and that is why various methods for improving search results, which can be implemented on the basis of expanding the original question, have been developed.

Our goal is to utilize knowledge of document similarity contained in textual databases to obtain a larger amount of relevant documents while minimizing those cancelled due to their irrelevance.

### 15.2 Issues with Metric Searching

The distance between the two documents $x$ and $y$ is the function $\delta(x,y) : X \times X \to R$ (where $X$ is a set of all documents), for which the following conditions apply:

\begin{align*}
\delta(x,x) &= 0 \quad (15.1) \\
\delta(x,y) &\geq 0 \quad (15.2) \\
\delta(x,y) &= \delta(y,x) \quad (15.3)
\end{align*}

Distance further requires the validity of triangle inequality. Triangle inequality is only valid when triad $x$, $y$ and $z$ abide by the following conditions:

\[ \delta(x,z) \leq \delta(x,y) + \delta(y,z) \quad (15.4) \]

Set $X$ and function $\delta$ create the metric space, which we identify as $(X, \delta)$.

#### 15.2.1 $\varepsilon$-ball and $\varepsilon$-$k$-ball

**Definition 15.1.** For given $x \in X$ and $\varepsilon \in \mathbb{R}^+$ (where $\mathbb{R}^+ = \{x \in \mathbb{R} | x \geq 0\}$), the set $B(x, \varepsilon) = \{y \in X; \delta(x, y) \leq \varepsilon\}$ is called the ball with the radius $\varepsilon$, or $\varepsilon$-ball centered at the point $x$.

**Definition 15.2.** For given $x \in X$, $\varepsilon \in \mathbb{R}^+$ and $k \in \mathbb{N}^+$, the set $B^k(x, \varepsilon) = \{y \in X; x = x_1, y = x_k, \sum_{l=1}^{k-1} \delta(x_l, x_{l+1}) \leq \varepsilon\}$ is called the $k$-ball with the radius $\varepsilon$, or $\varepsilon$-$k$-ball centered at the point $x$.  