Multi-dimensional Data Model of Textual Information
PAVEL ČECH
Department of Information Technologies
University of Hradec Králove
Rokitanského 62, Hradec Králové, 500 03
CZECH REPUBLIC
pavel.cech@uhk.cz

Abstract: - The key to understanding the meaning is in the proper representation of the textual information. The tree like representation of a text is not very convenient for large scale processing since the information about types of phrases, part of speech tags and actual terms are scattered in various tree nodes. The aim of the article is to propose a multidimensional model of textual data that would contain all the lexical information in a structure that enables further semantic processing. The approach in this article is based on the data modeling principles used in data warehousing.

Key-Words: - Multi-dimensional data model, data warehousing, star schema, text representation, lexical parsing.

1 Introduction
Information retrieval and text mining have already attracted attention of many researchers. The major challenge in these areas is the understanding of meaning reposed in the text. The key to understanding the meaning is in the proper representation of the textual information.

Today, it is known that the traditional term based representation of documents disregards the semantic differences contained in connections between the terms [29]. The novel concept based approaches utilize lexical parsing to extract semantics. Lexical parsing usually results in a tree like representation of a text (such as sentence). Such structure is however not very convenient for further large scale processing. The information about types of phrases, part of speech tags and actual terms are scattered in various levels of the tree depending on the text complexity.

The aim of the article is to propose a multidimensional model of textual data that would contain all the lexical information in a structure that enables further semantic processing. The approach in this article is based on the data modeling principles used in data warehousing [9, 10, 12].

The supposed major advantage of the proposed approach is in a relational calculus and standardized query language for fast and flexible retrieving of lexical information. Hence, suggested approach would enable easy exploration of the text using the various information retrieval and text mining techniques such as term proximity calculation, n-gram determination, Markov model application and Formal Concept Analysis.

The approach also addresses a problem stated by Castells et al that once the replacement of documents by “bag of knowledge” inevitably implies a loss of information [5]. Using this approach the documents can be fully reconstructed and thus new perspectives of the text could be reflected.

The paper presents the description of the proposed multidimensional model and a query possibilities as well as the verification of the approach on a sample set of approx. 200 documents.

2 Related work
There are various data models being used in the field of information retrieval and text mining. Basically, the two broad categories can be identified. The first category is term-based in which the vector-space model is being used [6, 17, 19, 20, 25]. The second category is concept or semantic based that utilizes ontology and a graph or tree like structures [5, 7, 11, 14, 16, 21]. There are also approaches that entail dimensionality or multi-dimensionality [27, 28, 29].

For example, Tseng proposed a multidimensional indexing structure, called D-tree [28], and corresponding query expressions MD2X (Multi-Dimensional Document Expressions) for document warehouses [27]. Tseng characterizes documents as a set of keywords and defines dimensions as a tree structure representing the hierarchical relations.
between keywords. The query expression MDX is an extension to MDX (Multi-Dimensional eXpressions) that is designed to include complete set of SQL-like constructs so that to enable the traditional OLAP (Online Analytical Processing) operation such as drill up and drill down on the above defined dimensions.

Similarly, Urbain et al designed a dimensional information retrieval model for genomics literature search. The model combines concept-based semantics and term statistics within multiple levels of document context in order to identify concise, variable length passages of text that answer a user query [29]. Their approach is based on a parallel dimensional indexing model in which the term index is used for identification of terms related to biological concepts and a concept index is used to search for resolved concepts.

Furtehr, de Mazière et al [6] analyzed 7,000 documents and organized them using special visualization techniques. In order to properly categorize the documents, they employed lexical parsing, n-gram finding and subsequent vectorization to use TF-IDF technique and clustering.

There also works that refer directly to data/web warehousing concepts. Thus, the issues of warehousing web data have been discussed in [1, 2, 3, 4] in which the concept of a Web Information Coupling and the WHOWEDA (Warehouse of Web Data) project have been introduced. The elaborated schema was a graph structure consisting of nodes representing web pages and links representing hyperlinks. They provided formal description of the schema and algebra with manipulation operators to query/navigate among interconnected pages.

Further, Fuller et al constructed an entity retrieval model for semi-structured information found in distributed and heterogeneous data sources such as world wide web or data of social networks [7]. In their approach they regard documents (HTML pages embedding RDF) to contain a set of statements describing one or more entities. The model use consists of entities and entity descriptions that form a graph-like structure in three levels.

It has to be noted, that none but one of the reviewed articles discuss the multidimensional model as it is being declared in the data warehousing concepts and principles. The multidimensionality refers rather to a vector space or the warehouse represents just the collection of data ignoring the subject orientation and other characteristics. Predominant data model is a graph like or tree like structure that are close to how ontologies are organized.

The proper application of data warehousing principles is found only in Urbain et al [29]. Urbain, however, does not refer to parts of the speech and phrases as dimensions and also the links between words are not considered.

### 3 Approach

The approach in this paper is based on the adoption of the data warehousing principles widely implemented in the so called business intelligence [15, 30]. Originally, data warehouse referred to a large volume data repository that enabled informed decision making of high level business managers. The term data warehouse was coined and promoted by Bill Inmon, who is also the author of commonly accepted definition of a data warehouse as “a subject oriented, nonvolatile, integrated, time variant collection of data in support of management's decisions” [10]. Correspondingly, the data warehousing can be regarded as a collection of methods and principles of how to build and operate data warehouse.

There are various approaches of how to realize the data warehouse using a relational database [10, 12, 13, 18, 24]. However, the core concept is a multidimensional data model or schema that is, in its simple form, often referred to as a star schema [9, 12]. The dimensions represent entities or perspectives through which the data should be viewed or analyzed. Dimensions are organized around particular measures (usually numerical data such as sales).

In relational databases the dimensions and measures are represented as tables (dimension tables and fact tables respectively). Unlike in traditional operational databases the data in data warehouse might not be normalized [9, 12].

The variant of the star schema, in which the dimensions are normalized, requires the data to be separated into several interrelated dimension that form a structure similar to a snowflake. This is used in cases in which data in dimensions can form a hierarchy (e.g. time dimension can be split into days, months, years or other time periods or location can be split into city, country, etc.). However, querying data stored in the snowflake schema is computationally more expensive since it entails more table joins.

In the application of a multidimensional model to textual data, the dimensions can be formed from the various lexical elements and structures (such as
words, sentences, documents and also parts of speech tags, etc. The measures contain occurrences of entities in dimensions. Thus the occurrence can refer to a particular occurrence of a certain word in a given sentence from a given document.

Populating the model would require some lexical parsing that might be achieved using the various tools and libraries such as Apache OpenNLP [26], Stanford NLP [23] or GATE [22].

3.1 Proposed multi-dimensional data model
The proposed model consists of the following dimensions (see fig. 1):

- **Words** – individual words; serves as a unique collection of all words indentified;
- **POS** – parts of the speech tags of words (such as noun, adjective, etc.); for example the Penn Treebank [8] tagset is suggested to be used;
- **Phrases** – phrases that the words are parts of (such as noun phrase or prepositional phrase); optional if no information about the phrase would be stored in that case the phrase would be stored in the fact table;
- **Sentences** – sentences as containing the words and are parts of the paragraphs and documents; optional if no information about the sentence would be stored in that case the sentence would be captured as order in document or paragraph in the fact table;
- **Paragraphs** – paragraphs as consisting of sentences; the paragraphs would serve for providing context and priority (e.g. paragraph with title might have greater weight than normal text). This dimension is optional if in some documents paragraphs might not be identified.
- **Documents** – documents such as pages are characterized by the title and location i.e. URL in case of pages or path in case of files in a document repository.
- **Word_Links** – links between consecutive words; the link is realized between two word occurrences; this dimension might not be necessary since the orders of the various parts are stored, thus when word occurrences are properly ordered the next word occurrence represents the next word. However, linking the words would allow easier manipulation using a relation calculus in which there are no operators for next or previous records. The **Documents**, **Paragraphs**, **Sentences** dimensions allow construction of a hierarchy, however, that would be more computationally demanding. Hence, the star schema is followed linking all dimensions directly to the fact table. Still, the dimensions are suggested to be normalized it means there will be no redundancy. In case of sentences and phrases are to be calculated a new fact tables for sentence or phrase occurrence might be created. However, the same information might be stored in the word occurrences table.

If **Word_Links** are normalized it would be characterized by link frequency that would be increased whenever the same consecutive pair of words is encountered. This enables pattern identification based e.g. on n-gram computation. In a simple form the model is suggested to contain one fact table called **Words_Occurrences** that would connect the data from the above mentioned dimensions. In relational databases the connection is realized using the foreign keys concept. Apart from the foreign keys the fact table would contain the **word_order** that is the order of a given word in a sentence.

3.2 Querying proposed data model
The proposed data model can be queried using the standardized SQL (Structured Query Language) or
MDX (Multi-Dimensional eXpressions). This would allow for easy manipulation with data since various views can be defined and also the off-the-shelf OLAP tools might be used to analyze the data. Thus, for example an OLAP cube might be constructed above the data to compute aggregates and to allow drill down and drill down operations.

The following SQL statement might be used to reconstruct the whole text enriched with part of speech tags and phrase types:

```sql
SELECT
    w.word,
    pos.tag,
    phr.type,
    wo.document_id,
    wo.word_order,
    wo.sentence_order
FROM
    dbo.tf_word_occurrences AS wo
    INNER JOIN dbo.td_words AS w
    ON wo.word_id = w.id
    INNER JOIN dbo.td_pos AS pos
    ON wo.pos_id = pos.id
    INNER JOIN dbo.td_phrases AS phr
    ON wo.phrase_id = phr.id
ORDER BY
    wo.document_id,
    wo.sentence_order,
    wo.word_order
```

There is also a possibility to identify word pairs. The following SQL statement would identify all adjective noun pairs:

```sql
SELECT
    v1.word AS adjective,
    v2.word AS noun,
FROM
    dbo.vx_wo_words_pos AS v1
    INNER JOIN dbo.vx_wo_words_pos AS v2
    ON v1.wo_after_id = v2.id
WHERE
    (v1.tag = 'JJ') AND (v2.tag = 'NN')
```

The pattern might be extended to include also adjectives that would further specify the concepts.

### 4 Verification

The verification of suggested approach has been conducted on the corpus of 180 documents or pages from Wikipedia. In that corpus over 40 thousands unique words were identified with more than 400 thousand word occurrences in 19 thousand sentences.

In order to analyze the documents, the suggested multidimensional model based on a star schema was created. Lexical parsing have been provided using OpenNLP framework [26]. During the lexical parsing the paragraphs were not identified and thus the Paragraphs dimension was omitted. Similarly, the sentence order in the document and word links between successive words were stored as a part of the word occurrence fact table.

The model enables the easy calculation of various aggregations. The Table 1 presents a sample of noun counts for selected documents.

<table>
<thead>
<tr>
<th>Document title</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolutionary programming</td>
<td>68</td>
</tr>
<tr>
<td>Candidate solution</td>
<td>76</td>
</tr>
<tr>
<td>Genetic representation</td>
<td>84</td>
</tr>
<tr>
<td>Chromosome (genetic algorithm)</td>
<td>89</td>
</tr>
<tr>
<td>Business intelligence tools</td>
<td>91</td>
</tr>
<tr>
<td>Social behavior</td>
<td>93</td>
</tr>
</tbody>
</table>

Table 1: Count of nouns in selected documents
More complicated aggregations are also possible. Using the above presented SQL statements the table 2 shows the frequency of keywords consisting of an adjective and noun in a document titled Computational science.

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>computational science</td>
<td>12</td>
</tr>
<tr>
<td>scientific computing</td>
<td>5</td>
</tr>
<tr>
<td>numerical analysis</td>
<td>3</td>
</tr>
<tr>
<td>computational engineering</td>
<td>2</td>
</tr>
<tr>
<td>scientific computation</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Frequency of adjective noun pairs

Computing the frequency of keywords can serve as a basis for the TF-IDF calculations and for other analysis.

5 Conclusion
The multi-dimensional data model appeared to provide fast and flexible way of retrieving lexical information. The query mechanism is based on the standardized query language and provides for further semantic analysis of the text.

The major disadvantage of the approach is the extent of the data. The possibility to fully reconstruct the analyzed text is obviously at the expense of the disk space. However, in space saving alternative mainly the word occurrence fact table would grow proportionally with the length of the text analyzed given that there is relatively small number of unique words that are used in a language.

References:


