

Word Taxonomy for On-line Visual Asset Management and Mining

Osmar R. Zaïane* Eli Hagen** Jiawei Han**

*Department of Computing Science, University of Alberta, Canada, zaiane@cs.uaberta.ca

**School of Computing Science, Simon Fraser University, Canada, {hagen, han}@cs.sfu.ca

Abstract

We have designed and implemented MultiMediaMiner, a system prototype to mine high-level multimedia information and knowledge from large multimedia repositories like the WWW. WordNet, a semantic network for English, was used to clean and transform sets of keywords extracted from Web pages to index multimedia objects contained in these pages. WordNet was also enriched and used to generate concept hierarchies necessary for interactive information retrieval and the construction of multi-dimensional data cubes for multimedia data mining with MultiMediaMiner.

1. Introduction

Data mining, the process of extracting implicit useful knowledge from large databases, has attracted the interest of many researchers in many fields. Even though data mining is well defined [6], the scope of mining and the range of information discovered are relative to the application and the type of data mined. Commonly, data mining tends to mine high-level information to extract rules for characterization, classification, clustering, association, etc. When using concept hierarchies it is possible to drill down along the hierarchy from a high level concept down to the raw data in the database. This process, if done intelligently, can also lead to interactive information retrieval.

The data mining we propose is on large multimedia databases. In our implementation we chose the Internet as the source of images and video sequences. Due to its unstructured and dynamic nature, the Internet is not well suited for conventional data mining techniques. In order to constrain the semi-structured data by a schema, we favoured building a database of metadata describing multimedia objects on the Internet and mining the multimedia metadata database. Keeping links to

the original images and videos allows us to go back to the resources and thus accomplish resource discovery along with knowledge discovery. Along with the visual clues extracted from the images, we wanted to automatically associate keywords to the images. These keywords were extracted from the web pages.

Research in data mining is flourishing and becoming widespread. Rapid progress is being reported in many areas and applications of data mining. However, despite the fact that multimedia has been the focus of many researchers for indexing, modeling, storing, etc., nothing substantial, in comparison to the relevance of the field, has been done in multimedia data mining, whereas there has been interesting research in text mining from textual documents [3,4] and Web or semi-structured data querying and mining [12,8,1,10,5].

Concept hierarchies are frequently used in data mining and are often the central data structure. However, these concept hierarchies are built manually, making the process tedious and domain limited. In our implementation, we do not claim to have fully automated the construction of the concept hierarchy, but our algorithm builds a hierarchy rich enough that can be improved by domain experts. Any knowledge discovery process involves data collection, data cleaning and transformation, data normalization and representation, the mining itself, and discovered knowledge representation and visualization. For lack of space, in this abstract we focus solely on the data cleaning and transformation and data representation of the keywords extracted.

2.System Overview

The *MultiMediaMiner* system [13,14] is a data mining system used to discover high-level information and knowledge from multimedia databases, like summarization of multimedia characteristics, classification of multimedia objects, and association between object features. It is a hybrid system borrowing technologies from *DBMiner* [7], a system for data mining in large relational databases and data warehouses, and *C-BIRD* [9], a system for Content-Based Image Retrieval from Digital libraries. While *DBMiner* applies multi-dimensional database structures, statistical data analysis, and machine learning approaches for mining different kinds of rules in relational databases and data warehouses, *C-BIRD* pre-processes images and videos to extract relevant features (colours, textures, layout, etc.), and matches queries with image and video features in the database it creates. Images and videos are extracted from Web pages and pre-processed to extract visual features. The Image Excavator [9], a web crawler especially built to traverse the Internet and extract images and videos, retrieves keywords from web pages that contain the images

and associates them with the images as “linguistic descriptions”. Instead of using all words in web pages as keywords, contextual information, like HTML tags in web pages, are used to derive these keywords. For example, image file name and path if it contains a word or recognizable words, ALT field in the IMG tag, HTML page title, HTML page headers, parent HTML page title, hyperlink to the image from parent HTML page, and neighbouring text before and after the image, META tag placed in the HEAD element of the HTML page, can disclose valuable keywords related to an image. Headers and titles in web pages are rarely complete phrases, and file names and directory paths (which are also used for keywords) are often truncated words, acronyms or created words (example Elvis2, Int8086, redcar, etc.). To solve the problem we decided to use a dictionary to identify legitimate words after simple transformation. The transformations concerns especially compound words like “blue angels”, and collated words like “/images/airplane” in a URL, or “fighter1.gif” in an image file name. After first transformation, a morphological analysis reduces words and verbs to their canonical form (base, infinitive, singular, etc.). The list of such transformed words is then cleaned to eliminate illicit or unrecognized words.

A concept hierarchy of keywords is a directed acyclic graph of concepts. An arc from a concept A to a concept B indicates that A subsumes B denoting that A is more general than B. For example “Boeing 747” is subsumed by “commercial airplane” which is subsumed by “airplane”. Building a concept hierarchy on keywords is not a trivial task. Most applications that use such a structure build them manually. To address both issues, cleaning keywords and building hierarchies automatically, we opted to use the on-line dictionary, thesaurus, and semantic network *WordNet* developed in the Linguistics department at the University of Princeton [11,2]. In a first stage, WordNet is used as a filter to eliminate unknown words. In a second stage, WordNet is consulted, as a partial order of words, to get subsumption relationships (i.e. child-parent) between words in order to build our concept hierarchy. Unfortunately, WordNet lacks some domain related words. In our first experiment, for instance, we were indexing airplane related web pages. Words like “Boeing 747” or “F-15” do not exist in WordNet and we had to enrich the semantic network in order to add valid words that were rejected in the first round. All terms that are filtered out and rejected by WordNet are checked by a domain expert and are added to a secondary semantic network that is anchored to WordNet. For example, “Boeing 747” which is not recognized in the first stage is added to the secondary hierarchy under a new concept “commercial airplane” which anchored as a descendent to the concept “airplane” in WordNet. The secondary (domain related) semantic network becomes an integral part of the enriched WordNet and is used in the keyword cleaning and concept hierarchy

building. Figure 2 shows the integration of WordNet in the design of a text cleaning and concept building module. In Figure 1, thumbnails of commercial airplanes pertaining to the aircraft manufacturer Boeing are displayed. This user interface from MultiMediaMiner also allows the selection of a multimedia data set to be mined. The hierarchy of keywords on the left of Figure 1 is a section of the concept hierarchy automatically generated by visiting some web sites containing aircraft images. The user selects images by browsing the concept hierarchy and choosing a keyword at a given conceptual level, and “zooms in” to a lower conceptual level or “zooms out” to a higher conceptual level before retrieving the relevant images as source of the data mining process. The user interface can also be used as an on-line resource discovery mechanism by drilling through to the web pages containing the images.

A major challenge is the mere size of a keyword dimension. Ordinarily, without counting the leaf nodes of a hierarchy, the size of concept hierarchies rarely exceeds 50 to 100 nodes in a data warehousing application, but eventually, a keyword hierarchy will reach several thousand nodes even for a restricted domain. It will eventually become impossible to deal with the keyword dimension in a traditional manner since the data cube becomes too sparse and too large to handle in main memory. New algorithms for handling the multi-dimensional data cube should be introduced. A second challenge is the multi-valued nature of the keyword domain. An image may have many keywords associated to it. This makes difficult to add the keyword dimension to the multi-dimensional data cube since the aggregate values in the data cube are based on one value per dimension per image. For our current implementation, we have left the keyword dimension outside the data cube, limiting our data mining possibilities to only the remaining image attributes. Keywords are only used to select images for analysis. We are working on a new data structure that would allow us to take advantage of the multi-valued nature of the keyword dimension inside the data cubes, and thus be able to include the keywords in our data mining and on-line analytical processing which would be very useful for visual asset management.

References

- [1] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. L. Wiener. The LOREL query language for semistructured data, 1997. <http://www.db.stanford.edu/~abitebou/pub/jod197.lorel96.ps>.
- [2] R. Beckwith, C. Fellbaum, D. Gross, K. Miller, G.A. Miller, and R. Teng. Five papers on WordNet. *Special Issue of Journal of Lexicography*, 3(4): 235-312, 1990. Also available from <ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps>.

