Data mining the web
Oscar Djupfeldt, Avraz Hirori, Christoffer Kullman
Group 5

Introduction

• Structure mining
• Content mining
• Usage mining
Problem

- There is a huge amount of information on the web
- The information available on the web is very diverse
- Information on the web is exists in almost all types of formats
- The web is semi-structured because of HTMLs nested structure
- The web is linked
- Lots of redundant information

Web Crawling

- Getting data about web pages
- Web search
- Done in two ways:
  - Uniformed graph search
  - Guided, “informed” search
Structure Mining

- Analyzes which places a web page points to and which places point to a web page
- To find relevant data about a web page
- Used for web search and social networks

Document Structure

- Look at the HTML- or XML-structure of a web page
- The structure can reveal which data in a document is relevant as well as providing a context to the information
Document Structure

What do we get from this?

• Tagging and indexing
  ▫ Expand main index.
  ▫ Small, separate indices for faster access.

• Relevance ranking
  ▫ Terms location affects ranking. Different tags = different weights.
  ▫ Good at natural relevance ranking.
Document Structure

Problems
• Invisible words
• Misleading text in titles, metatags, etc.
• Missing or incorrect information

Anchor text
• Most significant in use due
• Additional information

Link Analysis - Hyperlinks

• Independent evaluation of web page popularity or authority
• Idea is from social networks
  ▫ Popularity
  ▫ Authority
  ▫ Prestige
• Ranking
• Link analysis algorithms:
  ▫ PageRank
  ▫ HITS
Link Analysis - PageRank

- Democratic – a links to b = a “votes” for b.
- The “voter” is also analyzed
- Designed for the “random clicking user” - likelihood a user will reach a certain page
- Iterative process

\[ PR(p_i) = \frac{1 - d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \]
Content Mining

- Focused on text mining
- Video
- Audio
- Images
- Structured records like tables

- Information retrieval

Information Retrieval

Information retrieval mainly uses two types of measurement, precision and recall.

- Precision: the proportion of correctly returned pages out of all the returned pages.
- Recall: of all the pages that are correct, how many are returned?
Information Retrieval

We typically have one of the following:
- Perfect recall and very low precision
- Perfect precision and very low recall

Content Mining

- Web search engines
- Web directories
Content Mining

• Classification
  ▫ Genetic algorithms
  ▫ Memory based reasoning or K-Nearest neighbors based classification (K-NN)
  ▫ Vector space model

Vector Space Model

• Page similarity
  \[ \cos(\theta) = \frac{v_1 \cdot v_2}{||v_1|| ||v_2||} \]

• Weighted vectors
  ▫ Term frequency-Inverse document frequency method
  \[ v_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T \]
  \[ w_{t,d} = tf_t \cdot \log \frac{|D|}{|\{t \in d\}|} \]
Clustering

- Used to further help the user find what they are looking for
- Refine searches on a more general keyword
Usage Mining

- Web logs
  File (or several files) automatically created and maintained by a server of activity performed by it.

  Typically added
  - Client IP address
  - request date/time
  - page requested
  - HTTP code
  - bytes served
  - user agent
  - referer

Usage Mining

Web Usage Mining consists of three phases
- Preprocessing
- Pattern discovery
- Pattern analysis
Preprocessing

- To clean up the data
- User identification
- Session identification
- Path completion

Pattern Discovery

- Clustering algorithm
- Dependency modeling
- Classification
- Association Rules
Pattern Analysis

- SQL
- Filter out uninteresting rules
- Visualization techniques
  - Graphic
  - Color

Usage Mining

- Example
Conclusions

- Useful for personalizing the Internet
- Good commercial use
- Simplifies searching and browsing
- Adds much needed structure to the data on the web
- Easy to manipulate
- No authority
- Needed

Bibliography