

COMP8400: Algorithms and Techniques for Data Mining

Web data mining

Lecture outline

- The Web as a data source
- Challenges the Web poses to data mining
- Types of Web data mining
 - Mining the Web page layout structure
 - Mining the Web's link structure
 - Mining multimedia data on the Web
 - Automatic classification of Web documents
 - Web usage mining
- Important information about paper presentation

The Web as a data source

- The biggest source of information
- Distributed, dynamic, linked, different data types
 - Web pages contain semi-structured data (HTML and XML), as well as free format text, images, videos, sounds, etc.
 - Many Web pages are dynamically created, often by accessing databases (for example online stores, information directories, search engines)
 - Web pages are linked
 - Some parts of the Web are only accessible to certain people (logins required)

Challenges the Web poses to data mining

- The size of the Web (billions of pages, Petabytes of data)
- The complexity of Web pages (more complex than any traditional databases or text document collections)
- The Web is a highly dynamic source of information
- The Web serves a broad diversity of user communities (with different backgrounds, culture, interests and knowledge)
- Only a very small proportion of the Web is truly relevant or useful
- Trust in Web site content is often questionable

Types of Web data mining

- Mining the Web page layout structure
- Mining the Web's link structure
- Mining multimedia data on the Web
- Automatic classification of Web documents
- Weblog mining

Mining the Web page layout structure

- Compared to plain text, a Web page is a two-dimensional presentation
- Rich visual effects created by different font types, formats, separators, blank areas, colors, pictures, etc
- Different parts of a page are not equally important

The screenshot shows a CNN.com International page with several annotations:

- Title:** CNN.com International
- H1:** IAEA: Iran had secret nuke agenda
- H3:** EXPLOSIONS ROCK BAGHDAD
- TEXT BODY (with position and font type):** The International Atomic Energy Agency has concluded that Iran has secretly produced small amounts of nuclear materials including low enriched uranium and plutonium that could be used to develop nuclear weapons according to a confidential report obtained by CNN...
- Hyperlink:**
 - URL: <http://www.cnn.com/>
 - Anchor Text: AI oaeda ...
- Image:**
 - URL: <http://www.cnn.com/image/>
 - Alt & Caption: Iran nuclear ...
- Anchor Text:** CNN Homepage News ...

Web page block— Better information unit

The screenshot shows a CNN.com International page with three importance levels assigned to different blocks:

- Importance = Low:** Points to the 'EXPLOSIONS ROCK BAGHDAD' headline.
- Importance = Med:** Points to the 'IAEA: Iran had secret nuke agenda' headline.
- Importance = High:** Points to the main text body of the IAEA article.

Web Page Blocks

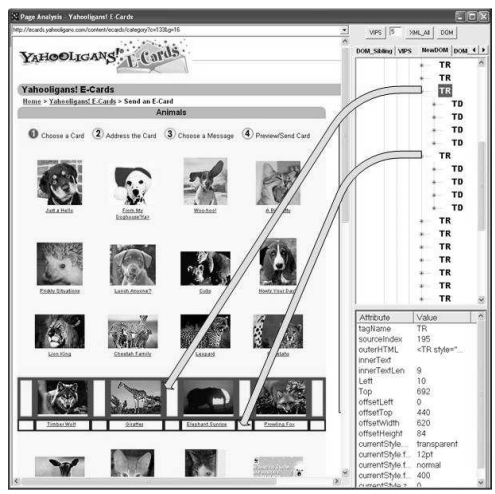
- Importance = Low
- Importance = Med
- Importance = High

Motivation for VIPS (Vision-based Page Segmentation)

- Problems of treating a Web page as an atomic unit
 - Web pages usually contain not only pure content, but also *noise* (navigation, decoration, interaction, etc.)
 - Web pages often contain multiple topics
 - Different parts of a page are not equally important
- Web pages have internal structure
 - Two-dimensional logical structure and visual layout presentation
 - > *Free text document* <
 - < *Structured document* >
 - DOM (Document Object Model) – tree structure of elements of a page
- Layout – the third dimension of Web pages
 - First dimension: content
 - Second dimension: hyperlink

Is DOM a good representation of page structure?

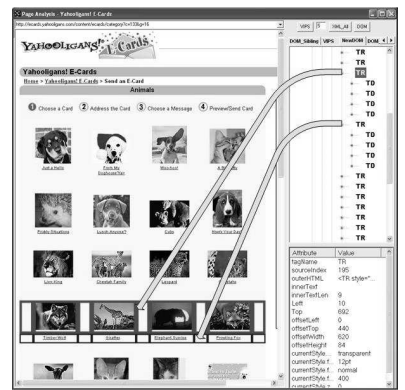
- Extract structural tags such as <P>, <TABLE>, , <TITLE>, <H1> to <H6>, etc.
- DOM is more related to content display, does not necessarily reflect semantic structure
- How about XML?
- A long way to go to replace HTML



VIPS algorithm

- Motivation
 - In many cases, topics can be distinguished with visual clues, such as position, distance, font, color, etc.
- Goal
 - Extract the semantic structure of a Web page based on its visual presentation
- Procedure
 - Top-down partition the Web page based on separators
- Result
 - A tree structure, each node in the tree corresponds to a block in the page
 - Each node will be assigned a value (degree of coherence) to indicate how coherent is the content in the block based on visual perception
 - Each block will be assigned an importance value
 - Segments obtained by VIPS are more semantically aggregated than in DOM tree

Example of Web page segmentation



DOM Structure



VIPS Structure

- Can be applied on Web image retrieval

Mining the Web's link structure

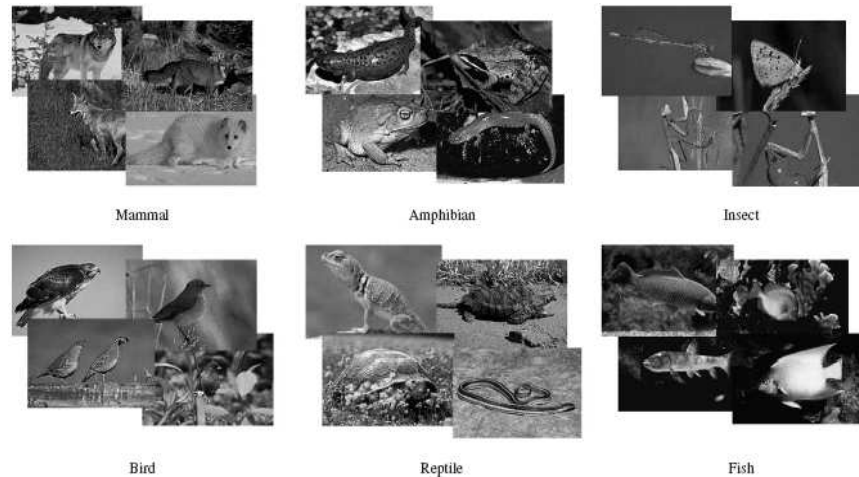
- The Web is a massive *graph* (Web pages are *nodes*, hyperlinks between them are *edges*)
 - Can use graph and link mining approaches (textbook chapter, not covered)
- Example: Find *authoritative* Web pages on a certain topic
 - A page many other pages point to
 - Not as easy, for example www.google.com does not explicitly contain "Web search engine"
 - Commercial and competitive interests, such as advertisements, distort the picture, as do many navigational links
 - To find authoritative Web pages, use *hub* pages (pages that provides links to many authoritative Web pages)
 - Example hub page: a personal home page with a list of recommended links
 - HITS (Hyperlink-Induced Topic Search): Start from search query, get *root* page set, which is then expanded, and a weight-propagation is iteratively conducted for hub and authoritative page weights

Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Mining multimedia data on the Web

- Data includes images, video, audio, graphs, etc.
- Mostly embedded into Web pages, often via hyperlinks
- Increasing demand for effective methods to organise and retrieve multimedia data
- Web page layout mining can be used to find multimedia blocks
- Example: Classify images
 - Use VIPS to identify multimedia blocks in a Web page
 - Use textual description around images for classification/categorisation
 - Use block-level link analysis (rather than page level link analysis such as Google's PageRank)

Example image categories



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Automatic classification of Web documents

- Supervised classification approach, training data required (classified Web pages)
- Classification done using keyword-based classification methods (text mining approach)
 - A Web page is seen as a document of terms
 - Take structure into account (block-based page content analysis)
 - Hyperlinks contain high-quality semantics (use to improve classification quality)
 - Hyperlinks surrounding a document (meta data, header, footers, etc.) are noise, can actually decrease classification accuracy
 - Often a very imbalanced problems, possibility to use one-class classifiers
- Large amount of research into *semantic Web*
 - Bring structure to Web based on semantic meaning of Web pages
 - Build *ontologies* of the Web (data model that represents a set of concepts)
 - Web page classification will help extracting semantic meaning of Web pages

Web usage mining

- Mining Web log records to discover user access patterns of Web pages
 - For example: "after looking a digital camera pages, 70% of users will look at memory card pages"
- Web log entry: URL requested, source IP address, time stamp, browsers details, cookies, etc.
 - Low level details, need to be cleaned, condensed, and transformed
 - Use data stream mining techniques
- Apply association and frequent pattern mining, and trend analysis
- Applications
 - e-Commerce, improve Web system design (navigation and caching), Web page pre-fetching, adaptive Web sites (depending upon user's history)

Student paper presentation

- I need papers and presentation preferences ASAP!
- Read specifications and what is expected carefully on COMP8400 Web site
- Draft session schedule made available on *Forum* soon – please check and e-mail me if you cannot make the session you are scheduled in!
- E-mail me your slides **as PDF file** (NOT Powerpoint format!!) at least two days before your presentation
- **DO NOT bring your laptop** – no time to switch

What now... things to do

- Read text book sections 10.4 and 10.5
- Work on student presentation and report (submit by Tuesday before your presentation!)
- Work on assignment 2 (due Wednesday 27 May 5 pm)
- Read papers for tutorial 4 (next week), will be made available online soon