Table of Contents
Introduction to Web Mining ................................................................. 4
    Web Mining ..................................................................................... 4
    Types of Web Mining ...................................................................... 5
Web Crawlers ....................................................................................... 6
    What is a Web Crawler ................................................................. 6
    Types of Web Crawlers ................................................................. 7
    Properties of Web Crawlers ......................................................... 7
Competitive Intelligence ................................................................. 8
    Introducing Competitive Intelligence ........................................... 8
    Kinds of information to look for ................................................... 8
    Places where information can be found ...................................... 8
Proposal ............................................................................................ 9
    Problem Definition ................................................................. 9
    Motivation .................................................................................. 9
    Objective: .................................................................................. 9
    Approach .................................................................................... 10
    References & Related Work ....................................................... 10
Introduction to Web Mining

Web Mining
Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data. Web mining uses many data mining techniques such as supervised learning (or classification), unsupervised learning (or clustering), association rule mining, and sequential pattern mining.

The Web has many unique characteristics, which make mining useful information and knowledge a fascinating and challenging task. Some of these characteristics are as follows:

- The amount of data/information on the Web is huge and still growing.
- Data of all types exist on the Web, e.g., structured tables, semi-structured Web pages, unstructured texts, and multimedia files (images, audios, and videos).
- Information on the Web is heterogeneous.
- Information on the Web is linked.
- The information on the Web is noisy.
- The Web is also about services.
- The Web is dynamic, as information on the Web changes constantly.
- The Web is a virtual society. People, organizations and automated systems interact thought the web.
Types of Web Mining

1. Web Usage Mining: refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests.

2. Web Content Mining: extracts or mines useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics. These tasks are similar to those in traditional data mining. However, we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc, for many purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer sentiments. These are not traditional data mining tasks.
3. Web Structure Mining: discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. For example, from the links, we can discover important Web pages, which, incidentally, is a key technology used in search engines. We can also discover communities of users who share common interests.

**Web Crawlers**

**What is a Web Crawler**
Web crawlers, also known as spiders or robots, are programs that automatically download web pages and analyze them looking for any referenced web pages. The pages discovered are in turn analyzed and the process is continued ad infinitum or until some stopping criteria is achieved. The pages discovered from the crawling process are usually treated as input to another system that further analyzes these pages for a variety of purposes including updating the indexes of search engines, email harvesting and website monitoring. Over the course of the paper, we will discuss crawler basics, the most common types of crawlers, the issues faced by contemporary crawlers and finally shed a light over some of the hot topics in research regarding crawlers.
Types of Web Crawlers

1. Universal Crawlers: Universal crawlers manage their frontiers as first-in-first-out (FIFO) queues. In this case, the crawlers would act as breadth-first crawlers. Pages to be analyzed are extracted from the head of the queue, while newly discovered pages are added to the tail of the queue.

2. Preferential Crawlers: Preferential crawlers manage their frontiers as priority queues. In this case, preferential crawlers act as best-first crawlers. As the crawler adds a newly discovered URL to the frontier, it is assigned a priority based either on the indegree of the target page, content properties regarding the target page, the proximity of keywords to the discovered URL in the source page or some other predefined measure. The URL with highest priority currently in the frontier is the one de-queued for crawling. As with universal crawlers, preferential crawlers are sensitive to the seed pages with which they are initialized.

Properties of Web Crawlers

1. Quality (Interesting Objects): Relevant to the user’s focus of interest.
2. Volume: To retrieve as many quality object as possible
3. Freshness: How Recent and up-to-date the retrieved objects are.
Competitive Intelligence

Introducing Competitive Intelligence.
Competitive intelligence refers to the process of gathering and analyzing information about products, domain constituents, customers, and competitors for the short term and long term planning needs of an organization.

Many major companies, such as Ernst & Young and General Motors, have formal and well-organized CI units that enable managers to make informed decisions about critical business matters such as investment, marketing, and strategic planning. Traditionally, CI relied upon published company reports and other kinds of printed information. In recent years, Internet has rapidly become an extremely good source of information about the competitive environment of companies and has been reported by a Futures Group survey in 1997 to be one of the top five sources for CI professionals.

Kinds of information to look for.
- Competitor Profiling (background, finance, marketing, personnel).
- Market products, product features and performance.
- New accounts, proposals, contracts and financial incidents.
- Customer preferences and opinions.
- New technologies, R&D and patents.

Places where information can be found.
- Competitor Web Sites
- Suppliers and Customers Web Sites
- News Web Sites
- Data Providers & Web Sites
- Community Articles and Blogs
- Social Networks
Proposal

Problem Definition
Dedicating human resources for harvesting relevant information from the World Wide Web has proven to be inefficient due to the vast size of information available and its distribution over numerous resources online.

Motivation
Delivering such a domain specific crawler that is able to collect useful separate pieces of information from around the web automatically and efficiently can reduce the cost of competitive intelligence process and give a better insight to the firm using such a tool.

Objective:

Objective is to develop a CI Spider that has the following capabilities:

1. Domain-aware spider: Continuously detecting slimier and/or related business entities to the domain.
2. Automatic Topic Tracking: This is the process of autonomously discovering emerging topics and gathering content that is relevant as it’s added.
3. Information Organization: Assembling the information gathered into topic based groups to facilitate their analysis by domain experts.
Approach

1. Make use of domain knowledge. This can be achieved by utilizing the concept of ontology and ontology-based search. This can make the spider domain-aware and focused on the entities of the specific domain that it was designed for.

2. Use meta-search for seed links. Make use of the result of famous search engines such as Google, Yahoo and MSN as seeds to the CI Spider to crunch the web looking for relevant information.

3. Integrate content-based and link-based similarity ranking. As a large amount of information is retrieved, these information should be ranked based on its quality (relevance to the topic). The ranking should base on both content and links related to the extracted page.

4. Relevance feedback. The CI Spider should be able to accept feedback from the user classifying whether the piece of information is related or not so it can adjust itself later.

5. Extract information from social networks. Mining social networks is a challenge by itself but it is very useful when trying to obtain a competitive edge gathering people opinions, suggestions, and comments about a product of your company or a competitor.

References & Related Work


