Automated Discovery of Brand Piracy on the Internet

Ulrich Pinsdorf and Peter Ebinger
Fraunhofer Institute for Computer Graphics
Fraunhoferstr. 5, 64283 Darmstadt, Germany
{ulrich.pinsdorf|peter.ebinger}@igd.fraunhofer.de

Abstract

The Internet has become a key instrument for establishing and promoting product brands. Companies are very interested in detecting any misuse of their logos, plagiarism, grey imports, and slander campaigns to protect the reputation of their brands. So-called Brand Protection *is very costly and requires a great expenditure of human resources especially when concerning electronic media.*

This article describes a software architecture that enables flexible observation and examination of Internet content regarding e.g. unauthorized use of logos, damage to brand names or shady sales offers. These algorithms check web content over several steps including a semantic analysis. The described system can be flexibly adjusted to specific monitoring tasks and inquiry orders to serve Internet detectives as an effective tool.

1. Introduction

The Internet has become an important instrument for the development and establishment of brands. Companies want to reinforce their presence on the Internet by actively governing and protecting the development of their brand capital. This requires monitoring the formation of opinions on the Internet, the use of brand names and their respective logos, and the distribution of knock-offs or grey imports on the Internet.

The U.S. jeweler store TIFFANY and other producers of luxury goods have monitored online auctioneer EBAY as well as others and sued them because imitations of luxury goods were offered there. TIFFANY specialists claim that about 73 percent of the articles offered on EBAY are fakes. The Wall Street Journal reports that the Internet has developed into the third largest market for fake luxury goods behind China and Italy.

The German chamber of industry and commerce (DIHK) estimates the annual economic damage caused by product and brand piracy at 29 billion Euro for Germany alone [7]. This figure indicates a loss of about 70,000 jobs.

Before it is possible to impose a sanction against brand pirates – such as caution, order of mandamus, claim for indemnification or complaint of an offense – it is necessary to collect facts to be able to confirm cases of suspicion. With regard to the Internet this method is called *Online Brand Monitoring*.

Specialized service providers offer brand and product monitoring for customers to survey and protect the use of their brands. This is achieved by so-called *Internet detectives* searching the World Wide Web for web sites that contain the brand name or the logo of their customer or that distribute fake merchandise via the Internet. The search is performed manually by using conventional search engines such as GOOGLE, YAHOO, etc. In general, the result of these inquiries is a detailed overview about brand violations including the individuals and companies involved. This information is given to the customer – in most cases the aggrieved party – who can then take further action. In certain cases the Internet detectives also do hidden test bargains to scrutinize the articles and to gather further evidence.

These investigations require the allocation of a high amount of human resources. Currently there are no specialized tools to support an Internet detective doing his work. The monitoring work is often monotonous and has a low hit rate. An increase of the efficiency of the search process is highly desirable, both for the customer and the provider. Monitoring and analysis tools should enable Internet detectives to gather web content efficiently and to analyze it according to customer-specific criteria.

This article proposes a flexible design that efficiently supports Internet detectives searching for brand piracy on the Internet. It describes a software architecture that enables flexible observation and examination of Internet content regarding e.g. unauthorized use of logos, damage to brand names or shady sales offers. The further text is structured as follows: the next section defines the requirements for the analysis tool; related work is described in section 3; section 4 presents the concept of our solution which is described in

more detail in section 5; finally section 6 summarizes and concludes the article.

2. Objectives and Requirements

The detection of brand piracy, grey imports or unauthorized use of brand logos requires a systematic semantic analysis of external web content. *External web content* is provided by portals, single web sites, Internet shops, and Internet auction platforms operated by third parties.

The following three model scenarios show the necessary bandwidth for some sample inquiries:

Scenario 1 The manufacturer of a product looks for offers of online stores with a price so low that with a certain probability they are grey imports or even fakes.

Scenario 2 A commercial provider of digital images on the Internet wants to find out which images appear on other web pages. For this reason he has marked them with a digital watermark.

Scenario 3 A certain company wants to find offers on an online auction that contain the brand logo of the company, although the vendor is not an authorized dealer.

The search criteria are – compared to conventional search engines – very "soft", so they can hardly be covered by simple algorithms. At the same time the search is very specific and adapted to the requirements and general conditions of the customer.

The fact that the search should be done "undercover" leads to further bounding conditions. The automatic search mechanism should not reveal itself with noticeable and traceable signatures. Therefore a highly *distributed architecture* is desirable, controlled and managed from a central workplace. Such a distributed system reflects regular user behavior and also provides scalability and load sharing.

The issues and requirements can be summarized as follows: a secure distributed system is needed that allows to semantically analyze external, publicly available web contents. The overall system shall work autonomously and be adaptable to new tasks. The job of the Internet detective is it to select cases that suggest potential brand piracy from the automatically generated list of discovered hints according to his own judgment.

3. Related Work

Significant research for the realization of so-called web crawlers has been accomplished in the area of *Information Retrieval* [14,15]. Search engines usually use a lexical analysis for archiving and the extraction of keywords from Internet contents [3, 6]. In general, content-based analysis is based on an evaluation function P that maps a content I and a request A to a scalar value $P(I, A)$.

The goal of *Focused Crawling* [5] is to specifically find the relevant pages for a defined set of subjects. These subjects are not specified by keywords but by sample documents. Instead of collecting and indexing all available web content to be able to answer any possible request only those references to other web pages are followed that are likely to have the highest relevance. In this way irrelevant regions of the Web are avoided. So Focused Crawling reduces the indexing effort of search engines since only web pages are examined that are deemed important.

An algorithm for *autonomous cooperative web crawlers* and a distribution protocol are presented in [10]. Cooperating web crawlers inform each other about modifications on the Internet to get a better and more up-to-date mapping of the World Wide Web than it is possible with traditional web crawlers. Cooperating web crawlers can share their results with other crawlers and mutually benefit from the results. The gain in performance that results from cooperation grows with the number of crawlers that join in the group.

In 1996 Etzioni coined the term *Web Mining* [8]. Content-related queries are known from *Data Mining* and *Image Retrieval*. Data Mining links unstructured data and derives new information that was not available in this way before and therefore provides an added value to the user. Kosala and Blockeel [9] give a good overview on web mining literature. Furthermore they describe in three categories of web mining technologies. The approach in this article fits in the category "Web content mining".

Systems for automatic image retrieval [12] analyze images according to the content that they represent. This way they provide an abstract view of the content independent of the particular digital representation which allows a fuzzy description of the requested images.

Tim Berners-Lee, one of the founders and masterminds of the Internet, suggests to semantically link web contents [1, 2]. This would replace the presently prevalent syntactic search and lead to a semantically linked and searchable information network, the so-called *Semantic Web*.

4. Proposed Solution

This section gives a overview of the proposed approach towards a solution for the described problem. The main requirements that were deduced in section 2 are that the system should be highly distributed, flexibly adaptable, and secure. A core problem is the understanding of the unknown web content, at least to the degree that a statement can be made in response to an inquiry. In order to allow a structured analysis of unstructured data we chose a multistage process. The HTML code is analyzed and filtered in three steps: syntactically, semantically, and logically.

Each of the particular filters of the three filter stages – one stage will typically contain more than one filter – is an independent module. A network of filter and analysis servers can be equipped with any number of filters. These can even be exchanged and integrated into a running system.

For a specific inquiry filters of the individual stages are dynamically connected. This combination and the associated parameterization is called *inquiry order*. Figure 1 shows different filter stages and how they cooperate.

5. Data Extraction and Aggregation

An important bounding condition is the usually covert execution of an inquiry. This implies that we assume non-cooperative data sources. Inquiries are based on publicly accessible web pages and web portals.

These web pages are usually dynamically generated on the according web servers. Therefore data stored in business databases for offers, customer relations, workflow management systems, etc. is linked by PHP, SHTML, or JSP and presented to the user in a graphical web front end. HTML files are transferred to the user by the protocols HTTP or HTTPS. In general, no direct access to the back-end databases is possible.

When the web pages are generated the underlying structure of the data is lost. Therefore, the presented data must be transformed back into a structured representation for the extraction and the analysis of relevant information.

The acquisition of data and the search for relevant results is structured into the following steps:

1. Selection and generation of web addresses
2. Automatic retrieval of web content
3. Syntactic analysis of the retrieved data
4. Semantic analysis of the retrieved data
5. Logical analysis of the retrieved data
6. Analysis, evaluation, and assessment of the relevant hits by an Internet detective

Figure 1 shows the schematic structure of the data flow and the configuration of the different filters. In the first step the web addresses of interest are generated and passed on to the query database. Then the web pages are retrieved from the Internet and stored in a database. Afterwards the data is syntactically and semantically analyzed and evaluated.

The analysis system has a modular architecture and can be flexibly extended by new filter algorithms. In each step, the data is analyzed in several modules in parallel concerning different aspects. The results are stored in the database where they are available for the next processing step. In the following, the six processing steps are described in more detail.

5.1. Selection and Generation of Web Addresses

In the first step the web addresses that will be investigated are selected or generated. These addresses can be generated according to different criteria. In the simplest case the web addresses are taken from an existing list. They can however also be generated dynamically by means of a search engine, an Internet portal, or a web directory.

A manually entered *list of web addresses* can consist of, e.g., URLs of web offers mentioned in a chat room. Alternatively the web pages can be acquired by a request to a *search engine*, e.g. GOOGLE, using certain keywords. In this way it is possible to extend the search horizon in a fuzzy but goal-oriented manner and casually come across new web pages containing plagiarism or brand violations. Furthermore, known *web portals* can be browsed targeting a certain issue, e.g. offers in a certain product category at EBAY. Another source for URLs are *web directories*, e.g. YAHOO, which classifies web pages according to particular topics. This allows to select all Internet addresses listed in a certain category.

In any case the result of this process step is a list of URLs that is stored in the query database. The generated web addresses are the starting points for the crawling process.

5.2. Syntactic Analysis

The content of the web addresses in the query database are requested via HTTP or HTTPS and the retrieved data is pre-filtered and formatted. The query module acts like a usual web browser.

If the structure of the web pages is known (e.g. the basic structure of product web pages on a certain portal), adapted *syntactic site description schemes* can be defined for the syntactic analysis. These allow to automatically extract only relevant information. Furthermore a generic standard scheme can be used to extract objects and information which are of general interest for inquiries. This specifically concerns the structure of the web page, images, multimedia files, and text.

The relevant parameters are linked together to allow a thorough analysis of web objects such as text or images. Examples for relevant parameters are:

- URL of the page where the object is located
- URL of the object
- Name of the object
- Date and time
- The object itself (respectively a reference to it)
- General features such as format, size, height, width
- Alternative text for the image
- The position within the page structure

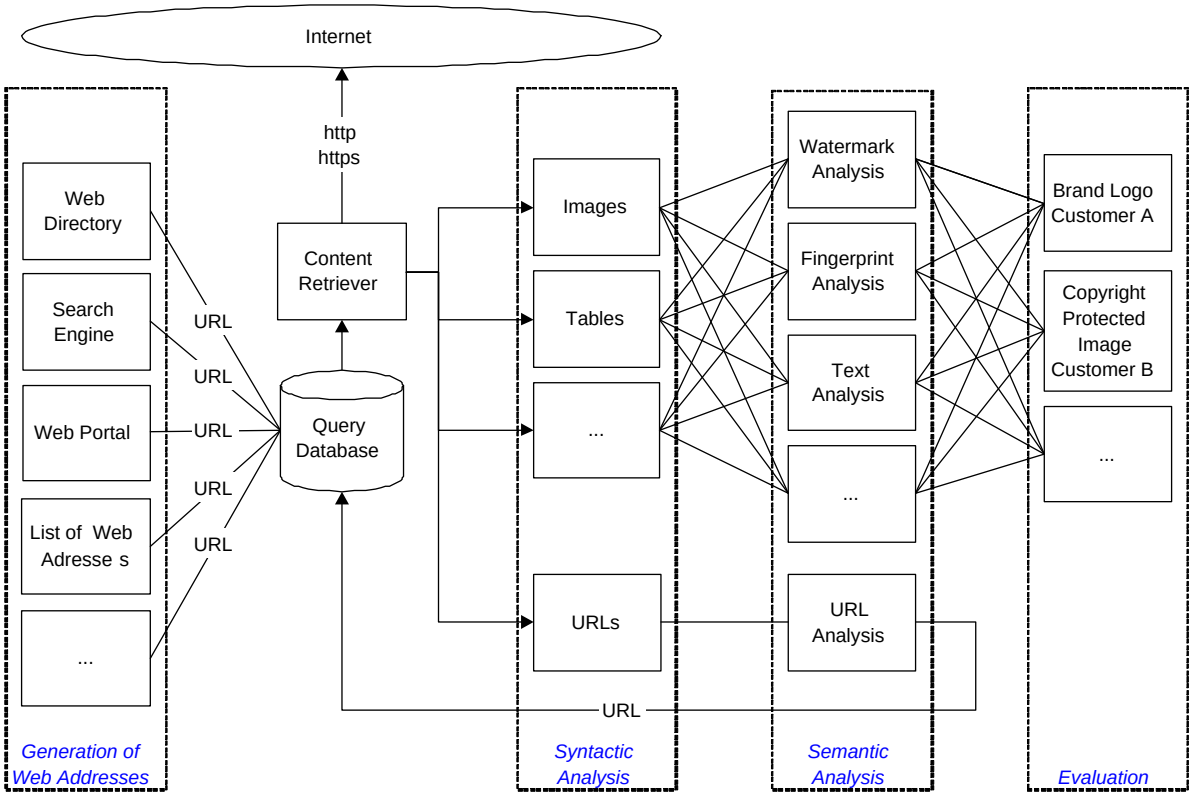

Figure 1. Dataflow within a filter and analysis server

5.3. Semantic Analysis

The next step is the semantic analysis of the data. This means that data elements can be set in relation to each other, analyzed and supplemented with additional information to represent their potential meaning. In a *semantic site description scheme* the meaning of the data elements (in the case of images e.g. dependent on position or image size) can be described. In this way single elements can be identified, e.g. as product logos. We use specific content-based algorithms for the semantic analysis which enable a more precise classification of the data elements based on the following aspects:

- Watermarking, e.g. for extraction of authorship
- Fingerprinting, e.g. for comparison to given samples
- File features, e.g. type, name, URL
- Text, e.g. product id and price
- Image features, e.g. size, resolution
- Meta information, e.g. image processing application
- Context reference, e.g. text content of the web page

Watermarking and fingerprinting are of special significance as base technologies for the semantic analysis. Watermarking allows users to hide information in multimedia content without remarkably affecting the quality. In the described application scenario mainly watermarking methods for images [4] are of interest since they allow companies to embed copyright information into company logos, product descriptions, or commercially sold image data.

Fingerprinting methods provide the opportunity to identify similar multimedia data by a small fingerprint independent of the particular digital representation. They can also be used if the contents already circulate on the Internet without watermark or if there are other reasons why it is not possible to modify the data. Here methods for images [16] are of special importance because images are essential components of web pages.

5.4. Logical Analysis, Evaluation

In the last step of the automated analysis process the data is analyzed with respect to the inquiry order and a report is generated. For this aim the different data elements of the semantic analysis are linked on the basis of a specific inquiry order and the results are compiled in a report. For example, the results of the text analysis regarding a product price, the watermark check for a product logo, and the verification of the examined URL are concentrated and evaluated, e.g. based on fuzzy logic, to estimate the probability of

a hit. Again, the results of the single analysis steps are written into the database.

The presented method is based on a flexible processing pipeline. The inquiry order of an Internet detective is reflected by the selection of specific processing modules and the corresponding data paths between them. This configuration must be easy to control and to adapt to new search strategies. From the point of view of the Internet detective the system looks like a common client/server based working environment.

Mobile agent technology is used to distribute specific algorithms to many different host systems in an efficient and flexible manner. For realization we chose the agent platform SEMOA [11, 13]. SEMOA is available as open source software, offers a component-based framework, provides code mobility, and also a number of sophisticated security mechanisms to protect both the inquiry agent and the distributed servers.

6. Conclusion

This article describes a system which enables an automated inquiry for brand piracy identification on the Internet. It shall support Internet detectives in tracing dubious web pages touching rights of brand owners. The proposed solution combines a number of advantages of existing technologies that are not present in a similar form within existing tools.

The proposed solution can be applied to arbitrary inquiries. The content-based search algorithms can be individually selected and combined at the user's discretion. The multistage analysis allows a flexible search depending on varying criteria based on the investigated web pages and the inquiry task. The distributed system of filter and analysis servers allows Internet detectives to do the inquiry of a web site by requests that originate from different servers and are performed at different times. Such inquiries are less noticeable than the query of a web crawler which constitutes an enormous progress in the brand protection domain.

Additionally, social control is an interesting subject considering brand piracy and copyright breach on the Internet. For example, the free online encyclopedia WIKIPEDIA trusts in their readers to detect copyright breaches contained in the included texts and images. If analysis tools such as the one described in this article were available free of charge to all Internet users, this would lead to a significant improvement of the contents and services available on the World Wide Web considering brand and copyright infringement based on the resulting social control.

Acknowledgment

Our research was co-funded by the German Federal Ministry for Education and Research within the SicAri project. As mobile agent platform we used the open source software SEMOA.

References

[1] T. Berners-Lee. Semantic web road map. Internal note, World Wide Web Consortium, Sept. 1998. See http://www.w3.org/DesignIssues/Semantic.html.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[4] S. Burgett, E. Koch, and J. Zhao. Copyright labeling of digitized image data. *IEEE Communications Magazine*, 36(3):94–100, March 1998.

[5] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specifi c web resource discovery. *Computer Networks*, 31(11–16):1623–1640, 1999.

[6] J. Cho, H. Garcia-Molina, and L. Page. Effi cient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30:161–172, 1998.

[7] *Keine Gnade für Plagiate. Gewerbliche Schutzrechte nutzen.* Deutscher Industrie- und Handelskammertag (DIHK), 2001.

[8] O. Etzioni. The world-wide web: Quagmire or gold mine? *Communications of the ACM*, 39(11):65–68, 1996.

[9] R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, 2(1):1–15, July 2000.

[10] G. L. McLearn. Autonomous cooperating web crawlers. Master's thesis, University of Waterloo, Canada, 2002.

[11] U. Pinsdorf and V. Roth. Mobile Agent Interoperability Patterns and Practice. In *Proceedings of Ninth IEEE International Conference and Workshop on the Engineering of Computer-Based Systems (ECBS 2002)*, 2002.

[12] V. Roth. *Sichere verteilte Indexierung und Suche von digitalen Bildern*. Ph.D. thesis, Technische Universität Darmstadt, Germany, June 2001.

[13] V. Roth and M. Jalali. Concepts and Architecture of a Security-centric Mobile Agent Server. In *Proc. Fifth International Symposium on Autonomous Decentralized Systems (ISADS 2001)*, Dallas, Texas, U.S.A., March 2001. IEEE Computer Society Press.

[14] G. Salton. *Information Retrieval: Data Structures and Algorithms*. Addison-Wesley, Reading, Massachussetts, 1989.

[15] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[16] J. S. Seo, J. Haitsma, T. Kalker, and C. D. Yoo. A robust image fi ngerprinting system using the radon transform. *Signal Processing: Image Communication*, 19, 2004.

Citation:

Pinsdorf, Ulrich; Ebinger, Peter:
Automated Discovery of Brand Piracy on the Internet.
In: Ma, Jianhua (Ed.) u.a.:International Conference on Parallel and Distributed Systems Workshops. Proceedings Volume 2 : ICPADS-2005 Workshops.Los Alamitos, Calif. : IEEE Computer Society, 2005, pp. 550-554