

UPGRADE is the European Journal for the Informatics Professional, published bimonthly at <http://www.upgrade-cepis.org/>



The European Journal for the Informatics Professional
<http://www.upgrade-cepis.org>

Vol. VIII, issue No. 1, February 2007

Publisher

UPGRADE is published on behalf of CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by **Novática** (<http://www.ati.es/novatica/>), journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*, <http://www.ati.es/>)

UPGRADE monographs are also published in Spanish (full version printed; summary, abstracts and some articles online) by **Novática**

UPGRADE was created in October 2000 by CEPIS and was first published by **Novática** and **INFORMATIK/INFORMATIQUE**, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>)

UPGRADE is the anchor point for UPENET (UPGRADE European NETWORK), the network of CEPIS member societies' publications, that currently includes the following ones:

- **Informatik-Spektrum**, journal published by Springer Verlag on behalf of the CEPIS societies GI, Germany, and SI, Switzerland
- **ITNOW**, magazine published by Oxford University Press on behalf of the British CEPIS society BCS
- **Mondo Digitale**, digital journal from the Italian CEPIS society AICA
- **Novática**, journal from the Spanish CEPIS society ATI
- **OCG Journal**, journal from the Austrian CEPIS society OCG
- **Pliroforiki**, journal from the Cyprus CEPIS society CCS
- **Pro Dialog**, journal from the Polish CEPIS society PTI-PIPS

Editorial Team

Chief Editor: Llorenç Pagés-Casas, Spain, pages@ati.es

Associate Editors:

François Louis Nicolet, Switzerland, nicolet@acm.org

Roberto Carniel, Italy, rcarniel@dgf.uniud.it

Zakaria Maamar, Arab Emirates, Zakaria.Maamar@zu.ac.ae

Soraya Kouadri Mostéfaoui, Switzerland,

soraya.kouadrimostefaoui@gmail.com

Rafael Fernández Calvo, Spain, rfcervo@ati.es

Editorial Board

Prof. Wolfried Stucky, CEPIS Former President

Prof. Nello Scarabottolo, CEPIS Vice President

Fernando Píera Gómez and

Llorenç Pagés-Casas, ATI (Spain)

François Louis Nicolet, SI (Switzerland)

Roberto Carniel, ALSI - Tecnoteca (Italy)

UPENET Advisory Board

Hermann Engesser (Informatik-Spektrum, Germany and Switzerland)

Brian Runciman (ITNOW, United Kingdom)

Franco Filippazzi (Mondo Digitale, Italy)

Llorenç Pagés-Casas (Novática, Spain)

Veith Risak (OCG Journal, Austria)

Panicos Masouras (Pliroforiki, Cyprus)

Andrzej Marciniak (Pro Dialog, Poland)

Rafael Fernández Calvo (Coordination)

English Language Editors: Mike Andersson, Richard Butchart, David Cash, Arthur Cook, Tracey Darch, Laura Davies, Nick Dunn, Rodney Fennemore, Hilary Green, Roger Harris, Michael Hird, Jim Holder, Alasdair MacLeod, Pat Moody, Adam David Moss, Phil Parkin, Brian Robson

Cover page designed by Concha Arias Pérez

"Gaia gateway" / © ATI 2007

Layout Design: François Louis Nicolet

Composition: Jorge Llácer-Gil de Rames

Editorial correspondence: Llorenç Pagés-Casas pages@ati.es

Advertising correspondence: novatica@ati.es

UPGRADE Newsletter available at

<http://www.upgrade-cepis.org/pages/editinfo.html#newsletter>

Copyright

© Novática 2007 (for the monograph)

© CEPIS 2007 (for the sections UPENET and CEPIS News)

All rights reserved under otherwise stated. Abstracting is permitted with credit to the source. For copying, reprint, or republication permission, contact the Editorial Team

The opinions expressed by the authors are their exclusive responsibility

ISSN 1684-5285

Monograph of next issue (April 2007)

**"Information Technologies
for Visually Impaired People"**

(The full schedule of UPGRADE
is available at our website)

Monograph: Next Generation Web Search

(published jointly with Novática*)

Guest Editors: *Ricardo Baeza-Yates, José-María Gómez-Hidalgo, and Paolo Boldi*

- 2 Presentation. The Future of Web Search — *Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo*
- 5 Efficient Sparse Linear System Solution of the PageRank Problem — *Gianna M. Del Corso, Antonio Gullì, and Francesco Romani*
- 12 Learning to Analyze Natural Language Texts — *Giuseppe Attardi*
- 19 SNAKET: A Personalized Search-result Clustering Engine — *Paolo Ferragina and Antonio Gullì*
- 27 The Multimodal Nature of the Web: New Trends in Information Access — *Luis-Alfonso Ureña-López, Manuel-Carlos Díaz-Galiano, Arturo Montejo-Raez, and M^a Teresa Martín-Valdivia*
- 33 Adversarial Information Retrieval in the Web — *Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo*
- 41 GERINDO: Managing and Retrieving Information in Large Document Collections — *Nivio Ziviani, Alberto H. F. Laender, Edleno Silva de Moura, Altigran Soares da Silva, Carlos A. Heuser, and Wagner Meira Jr.*
- 49 Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web — *Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras*
- 57 Yahoo! Research Barcelona: Web Retrieval and Mining — *The Yahoo! Research Team*

UPENET (UPGRADE European NETWORK)

- 59 From **Novática** (ATI, Spain)
Informatics Profession
The Maturity of IT Professionalism in Europe — *Sean Brady*
- 68 From **Pro Dialog** (PTI-PIPS, Poland)
Graphical Interfaces
Portable Declarative Format for Specifying Graphical User Interfaces — *Zbigniew Fryźlewicz and Rafał Gierusz*
- 75 From **Novática** (ATI, Spain)
Next-generation Web
Blogs: On the Cutting Edge of the Next-generation Web — *Antonio Miguel Fumero-Reverón and Fernando Sáez-Vacas*

CEPIS NEWS

- 83 Harmonise Project: Building up to the Final Report—*François-Philippe Dragnet*
- 84 News & Events: European Funded Projects and News Updates

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>.

Adversarial Information Retrieval in the Web

Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo

The Web is the killer application of the Internet. Without doubt, such a useful application is destined to be the subject of abuse, as others like e-mail are. Spam has invaded the Search Engines, the Social Networks, and moreover, the Web is also abused by its users and not only the content providers. Adversarial Information Retrieval (AIR) deals with the classification of content (or use of content) regarding its abuse quality, and faces an adversary (the abuser), who is ever trying to mislead the classifier. Search Engine spam detection, Web content filtering, and others, are instances of AIR in the Web. In this work, we review a number of AIR problems in the Web, along with some proposed solutions. We pay special attention to link-based Search Engine spam detection, and to Web content filtering, as representatives of a range of proposed techniques to reach high effectiveness in controlling Web related abuse.

Keywords: Adversarial Information Retrieval, Link Analysis, PageRank, Search Engine Spam, Web Spam, Web Filtering.

1 Introduction

As the amount of information and usage of the Web increases, so does its economic value and the interest in abusing it. Since search engines are the most prominent entry point to the Web, they are the focus of complex attacks named Search Engine Spam. Other forms of abuse include surfing inappropriate Web contents in schools, libraries and

the work place. What these kinds of abuse have in common is that they both can be addressed as text classification tasks in a so-called Adversarial Information Retrieval (AIR) setting [7].

The most prominent adversarial classification task is spam e-mail filtering [12]. Spam or junk e-mail messages advertising porn websites, Viagra pills, or asking for users' bank data, are sent in hundreds of millions each day, causing irritation, waste of time and severe economic damage. Spam filtering is a text classification task in which e-mail messages are classified as spam or legitimate. In spam fil-

Authors

Ricardo Baeza-Yates is Director of the new Yahoo! Research laboratories in Barcelona and Latin America (Santiago, Chile). Before that he was professor and director of the Center for Web Research at the Computer Science department of the University of Chile, and also ICREA (*Institució Catalana de Recerca i Estudis Avançats*) Professor at the Department of Technology of the *Universitat Pompeu Fabra* in Barcelona, Spain. He holds a Ph.D. in Computer Science from the University of Waterloo, Canada. He is co-author of the book *Modern Information Retrieval*, published in 1999 by Addison-Wesley, as well as co-author of the 2nd edition of the *Handbook of Algorithms and Data Structures*, Addison-Wesley, 1991; and co-editor of *Information Retrieval: Algorithms and Data Structures*, Prentice-Hall, 1992. Among other awards, he received the Organization of American States award for young researchers in exact sciences (1993). In 2003 he was the first computer scientist to be elected to the Chilean Academy of Sciences.

Paolo Boldi obtained his Ph.D. in Computer Science at the University of Milano, where he is currently Associate Professor at the *Dipartimento di Scienze dell'Informazione*. His research interests touched many different topics in theoretical and applied computer science, such as: domain theory, non-classical computability theory, distributed computability, anonymous networks, sense of direction, self-stabilizing systems. More recently, his works focused on problems related to the World-

Wide Web, a field where his research has also produced software packages used by many people working in the same area. In particular, he contributed to write a highly efficient full-text IR engine (MG4J), and a graph compression tool (WebGraph) that is state-of-art as far as compression ratio is concerned.

José-María Gómez-Hidalgo holds a Ph.D. in Mathematics, and has been a lecturer and researcher at the *Universidad Complutense de Madrid* (UCM) and the *Universidad Europea de Madrid* (UEM), for 10 years, where he is currently the Head of the Department of Computer Science. His main research interests include Natural Language Processing (NLP) and Machine Learning (ML), with applications in Information Access in newspapers and biomedicine, and Adversarial Information Retrieval with applications in spam filtering and pornography detection on the Web. He has taken part in around 10 research projects, heading some of them. José María has co-authored a number of research papers in the topics above, which can be accessed at his home page <<http://www.esi.uem.es/~jmgomez/>>. He is Program Committee member for CEAS (*Conference on Email and Anti-Spam*) 2007, the Spam Symposium 2007 and other conferences, and he has reviewed papers for JASIST (*Journal of the American Society for Information Science and Technology*), ECIR (*European Conference on Information Retrieval*), and others. He has also reviewed research project proposals for the European Commission.

tering, the content provider (the *spammer*) abuses the medium (the e-mail) by sending the consumer (the user) undesired messages, specifically prepared for avoiding detection by the user's anti-spam filter.

What makes the problem challenging is the fact that spammers are highly committed to avoid detection, because reaching more and more users is the way to get their economic benefit. One in one hundred thousand users buying a box of false Viagra is enough to get an important economic profit. The main property of Adversarial Information Retrieval tasks is the existence of an individual, the adversary, who is constantly improving his or her methods to make the system mislead, making wrong decisions. In fact, there is an endless war between the abusers and the classification system developers, with no expected winner at the end.

In this paper, we review a number of AIR problems in the Web, and focus on two of them, because they provide a good overview of the techniques being used for the whole range of problems. The problems we discuss in detail are link-based Search Engine Spam [13], and Web content filtering [15].

2 Adversarial Information Retrieval Problems on the Web

Spam is the paradigmatic problem in Adversarial Classification and many tasks have borrowed that word from the task of spam e-mail filtering. What they have in common is the existence of an abuser or adversary, who tries to make the classifier fail. While most of the time, the adversary is the content provider (i.e. the spammer who writes his or her e-mails or the Web master who tries to make his Web site get undeserved high rank in many Web searches, etc.), sometimes the user also tries to abuse the classifier (as children do when trying to access inappropriate Web content in schools, like pornography).

The goal of this section is to present an outline of the most relevant AIR tasks in the Web, and pointers to proposed solutions in the literature (if any). The tasks we discuss are organized in three groups: content spam, link spam, and usage spam. In the last section, we discuss emerging forms of AIR in the Web.

2.1 Content Spam

Content spam deals with the specific preparation of the content of a Web page, in order to make it misleading for search engines, and thereby get an undeservedly high rank for popular user queries. Also, we consider Web navigation abuse and, in consequence, Web filtering, a kind of Web content spam.

2.1.1 Content-based Search Engine Spam

Search Engines still have to return Web sites according to keywords entered by their users. The content providers can increase their impact by inserting popular keywords in their Web pages [13]. For instance, an electronic commerce Web site owner may put the most popular keywords according to public ranks like Google's Zeitgeist, even if they are not related to its business. The goal is getting his or her

Web site returned as a hit for a popular search keyword, attracting consumers to it. This form of abuse is often named "*keyword spamming*".

The abuser (the content provider) tries to get a high rank for these popular keywords by putting them in the title, URL or anchor HTML tags, making the keywords very frequent in his or her Web page, and even copying large pieces of text from popular Web sites. The keywords are often hidden by showing them in the same colour as the background of the Web page.

Ntoulas et al. [20] have investigated the text and format properties of keyword spamming pages. For instance, they have found a clear correlation between the "spamnicity" of a page, and the number of keywords in the page (the more words, the higher probability of being spam). Other useful features for detecting text-based spam are the number of words in the page title, the average word length, the amount of anchor text or the fraction of visible content.

2.1.2 Cloaking

Cloaking is a form of information disguising in which content providers serve different versions of a Web page to the users and to the Search Engine indexing robots [25]. This way, an unrelated Web page may be retrieved for a suitable search query.

An obvious way to address cloaking is to compare the versions of the Web page retrieved by different agents (e.g. the official Google or Yahoo! bots, and a bot disguised as a Firefox user). However, this method is limited by the ever-changing nature of dynamic Web pages: Web servers return pages with different news items, advertisements, etc. Wu and Davison [25] propose to compare different copies of the Web page obtained by two crawlers and two browsers. If crawler Web page versions are similar, and different to browsers' ones, which are in turn similar, the Web page is a good candidate of cloaking. As downloading four versions of a Web page is quite inefficient for a Search Engine crawler, the method is refined in [26] by filtering out those pages with no significant difference between one crawler and one browser versions.

2.1.3 Sensitive Content and Abuse

It is clear that some kinds of Web information are inappropriate depending on the place where they are accessed. For instance, casino games are not suitable for corporation employees at the work place, or pornography should not be accessed by children. In order to avoid this kind of abuse, with a clear economic impact (in terms of working time and bandwidth wasted, for instance), filters and monitors have been designed and are present nowadays in an increasing number of institutions. We discuss the techniques used by these systems in the Section 3.

2.1.4 Blog Spam

Web logs or just blogs may well be the most popular Web 2.0 tool, enabling millions of Internet users to have a voice in the Web, and leading to social networks which are

extremely useful for targeting advertisements. Because creating, editing and commenting blogs is so easy, they have been specially abused, in order to drive rank to even non-blog Web pages, through links in massively posted comments, and other tactics [16].

Blogs are a specific genre in the Web, with specialized Search Engines like Technorati. So, specific techniques for detecting different forms of blog spam (often named "splog") have been proposed. For instance, Kolari et al. [17] have proposed a Machine Learning approach based on text features for detecting spam blogs; this approach is very similar to some of those we discuss in the Sections 3 and 4. Mishne et al. [19] focus on blog comments, presenting an approach for detecting link spam common in blog comments by comparing the language models used in the blog post, the comment, and pages linked by the comments. In contrast to other link spam filtering approaches, their method requires no training, no hard-coded rule sets, and no knowledge of complete-web connectivity.

2.2 Link Spam

Perhaps one of the main reasons for the success of Google is the ability of its designers to recognize the social nature of the Web, that is, the Web as a place to build human relations. The most reliable social information in the Web is represented by hyperlinks, through which content providers demonstrate their interest and trust on others' contents. Today, most Search Engines make use of some form of link-based quality metric for Web sites, with direct impact on the rankings they return to users' queries. The most popular one is that used by Google, the PageRank [21]. At the end, these metrics are no more than variations of prestige measures used for scientific literature, based on citation graphs.

As scoring high in Search Engines drives users to your Web site, it has big economic value, and in consequence, some content providers have explicit interest on abusing link-based ranks. The straightforward method to improve the rank of a page is to build a network of highly connected Web pages (a *link farm*) pointing to it, with the hope of these sites propagating their rank through the network, to the target Web site. In the Section 4 we provide a detailed description of the methods used to detect and avoid this kind of abuse.

2.3 Usage Spam

Another important measure used by search engines is user clicks. Page clicks after a query may indicate good pages. Although page clicks are biased by the ranking and the interface of the search engine, their real distribution can be approximated by reducing this bias. However clicks can also be spammed. How we distinguish a click by a real user from another click by a software agent? Until the current time there have been very few public results for this problem (e.g. detecting recurrent IP addresses or strange click patterns).

This is a very difficult problem currently under research, in particular because this problem is related to an even more important click spam: false advertising clicks. In this case

each click implies a real monetary cost, so click spam must be detected to charge each advertiser only for clicks done by real people.

3 Case Study: Web Content Filtering

A very relevant form of abuse related to content is the access to inappropriate Web content in the workplace and in schools. The content analysis techniques used in Web filters offer a good representation of content-based adversarial classification, and we discuss them below.

3.1 The Problem of Accessing Inappropriate Web Content

The self-regulating nature of Web publishing, along with the ease of making information available on the Web, has allowed some content providers make offensive, harmful or even illegal contents available in Web sites across the world. This fact makes the use of filtering and monitoring systems a necessity in educational environments, and in the work place, to protect children and prevent Internet abuse.

In contrast to the previous tasks, the adversaries in this task are both the content providers and the content users. On the one hand, some content providers put illegal and harmful content in the Web, like hate speech (xenophobia and racism), conveniently disguised as legitimate political opinions. On the other, consumers make an inappropriate use of the Internet resource at the workplace (for instance, accessing casino games) or in schools and libraries (for example, accessing videogames or pornographic Web sites). While education and a suitable Internet Usage Policy are required to prevent this abuse, Web filters and monitors can be used to detect or prevent anomalous behaviour.

3.2 Techniques for Web Filtering and Monitoring

There are a number of filtering solutions available in the market, including commercial products like CyberPatrol or NetNanny, and open source systems like SquidGuard or DansGuardian. According to in-depth evaluations of these products (e.g. the one performed in the European Project NetProtect [6]), their filtering effectiveness is limited by the use of simple techniques, like URL blocking, or keyword matching. There is a need of more sophisticated and intelligent approaches to increase the effectiveness of filtering solutions. In this section, we describe the most popular approaches used for Web filtering, and focus on the most promising one, namely intelligent content processing.

The Web content filtering approaches can be classified in four major groups [18]:

- Self or third-party ratings, especially the use of Platform for Internet Content Selection (PICS) or Internet Content Rating Association (ICRA) ratings. Authors or reviewers label Web pages according to several types of content and levels, which are used by the filtering system to allow or block the pages according to the settings defined by the user or administrator. Unfortunately, only a small fraction of Web pages are labelled, and authors can inadvertently (or intentionally) label their pages with incorrect tags.

- Uniform Resource Locator (URL) listing, that is, maintaining a list of blocked and/or allowed web sites. A Web page is blocked if its URL contains a blocked URL, or its outgoing links point to blocked URL addresses. These kinds of lists can be automatically or manually built, but they are difficult to keep updated, and do not account for domain aliasing.

- Keyword matching, in which a set of indicative keywords or key-phrases (e.g. sex, free pics) are manually or automatically derived, form a set of Web pages to be blocked (e.g. containing pornography). A Web page is blocked if the number or frequency of keywords occurring in it, exceeds a predetermined threshold. This approach is prone to over-blocking, that is, blocking safe Web pages in which these keywords occur (e.g. sexual health, etc.).

- Intelligent content analysis, which involves a deeper understanding of the semantics of text and other media items (especially pictures), by using linguistic analysis, machine learning, and image processing techniques. The heavy cost of building linguistic analyzers and image processing components, their domain dependence (e.g. techniques for detecting nudes are quite different to those for recognizing Nazi symbols), and the delay caused by in-depth analysis, limit the applicability of these techniques.

The first three approaches, widely used in current filtering solutions, have proved quite ineffective, and have serious limitations [6]. We argue that intelligent content analysis is feasible, as far as the system design deals with delay issues, and linguistic and image processing are kept as shallow as possible. For instance, the POESIA project [15] is designed to have two levels of filtering: light filtering, for those Web pages that are not suspicious, or clearly inappropriate (e.g. pornographic); and heavy filtering, for those Web pages in which light filters are not able to give a clear judgment.

Linguistic and image processing techniques in the light filters are very limited and efficient, while heavy filters use more advanced (but shallow, anyway) methods, giving a more accurate but delayed answer.

As an instance of an intelligent analysis method, we next describe an approach to Web filtering as Automated Text Categorization, used in the Spanish pornography text filter developed in the POESIA project.

3.3 Web Content Filtering as Text Categorization

Automated Text Categorization (ATC) is the automatic assignment of text documents to predefined categories. Text documents are usually news items, scientific reports, e-mail messages, Web pages, and so on. Categories are often thematic, and include library classifications (e.g. the Medical Subject Headings), keywords in Digital Libraries, personal e-mail folders, Web directory categories (like Yahoo!'s), etc. Automated Text Categorizers can be built by hand (e.g. by writing rules for e-mail message filing in personal folders), or they may be constructed automatically, by training a text classifier on a set of manually labelled documents. This latter, learning-based, approach has become dominant, and

current techniques allow building ATC systems as accurate as human experts in a domain [23].

Pornography detection has been approached as learning-based ATC in several recent works, including e.g. [9] [10] [15] [18]. From these works, we can model pornography detection as a 2-class learning problem: train a classifier which decides whether a Web page is pornographic or not. In the learning phase, given two sets of pornographic (P) and safe (S) Web pages (the training collection), the following steps are given:

1. Each Web page in P or S is processed to extract the text it includes (pieces of text inside TITLE, H1, P, META; or, just all tags are stripped out), defining its text content. The text is tokenized into words, which may be stemmed and/or stop listed (ignoring those occurring in a function word list), producing a list of text tokens.

2. Optionally, a number of tokens is selected according to a quality metric like Information Gain. This step allows reducing the dimensionality of the problem, speeding up learning and may even increasing accuracy. The resulting set of tokens is the final lexicon.

3. Each page is represented as a term-weight (or attribute-value) vector, being the term's previous text token. The weights can be binary (a token occurs in the Web page, or not), Term Frequency (number of times the token occurs in the page), TF.IDF (the previous one times the Inverse Document Frequency), Relative Term Frequency, etc. [23].

4. Finally, a classifier is induced using a training algorithm over the set of training vectors and its associated class labels. Algorithms used in this problem include the probabilistic Naive Bayes algorithm and Bayesian Networks [9], variants of lazy learning [10], semi-supervised Neural Networks [18], and linear Support Vector Machines [15].

The first and third steps define the text representation model, which in this case is often called the 'bag of words' model. It corresponds to the traditional Vector Space Model in Information Retrieval. The classification phase involves, given a new Web page whose class is not known, its representation as a term-weight vector similar to those in the training collection, and its classification according the model generated in the learning phase.

This phase must be extremely efficient, avoiding long delays in Web page delivery when they are allowed (classified as safe).

Heppele et al. [15] have roughly followed this approach for building text-based pornography filters for English and Spanish, reaching very good levels of effectiveness. For instance, the English filter is able to detect around the 95% of pornographic Web pages.

4 Case Study: Link-based Search Engine Spam

As a representative of link-analysis techniques in the Web, we have focused in this section on how Search Engines take advantage of the connected nature of Web information, in order to improve results ranking, and how this kind of ranking is being abused. We outline the most promising solutions to this kind of spam.

4.1 Link-based Ranking

The idea behind link-based ranking is that the relevance of a page with respect to a given query cannot be determined only on the bases of the page content (textual or otherwise), but also on the hyperlink structure of the page and its vicinity, or even of the whole web. The advantage of this form of ranking over the traditional, content-based methods is that they tend to provide a more exogenous measure of relevance, which, besides being more trustworthy, also appears to be more resistant to spam.

Link-based ranking techniques may be further classified into dynamic and static; the former rank documents with respect to a given query (and thus can only be computed at query-time) whereas the latter are query-independent, and they can be rather interpreted as a measure of absolute importance of a web document. Even if the recent literature on the subject discusses both kinds of rankings, the real massive usage of dynamic algorithms is still unfeasible in the large scale, whereas static techniques are largely popular, and most today's search engines are believed to adopt them in one form or another.

The simplest static ranking is the so-called in-degree: a page is considered important if and only if it has many in-links, i.e., if there are many pages (called *in-neighbors*) that have a hyperlink towards it. Indeed, in-degree is a trivial measure of popularity; it is not really difficult to spam it, though, because one can just create many in-neighbours for a page with the only aim of deceiving the ranking algorithm, letting it believe that is more important than it should be.

A generalization of this idea is the well-known PageRank algorithm [21]. A suggestive metaphor to describe the idea behind PageRank is the following: consider an iterative process where every web page has a certain amount of money that will at the end of the process be proportional to its importance. Initially, all pages are given the same amount of money. Then, at each step, every page gives away all of its money to the pages it points to, distributing it equally among them: this corresponds to the interpretation of links as a way to confer importance. This idea has a limit, however, because there might exist groups of pages that "suck away" money from the system without ever returning it back. Since we want to disallow the creation of such oligopolies, we force every page to give a fixed fraction of its money to the State; the money collected this way is then redistributed among all the pages.

One can prove that this process indeed reaches a stable behavior for every taxation rate between 0 and 1 (but strictly less than 1); the parameter is called *damping factor* and it is a measure of the degree of importance we are giving to web links (in particular, when every page has the same PageRank). For traditional (and, in part, still mysterious) reasons, 0.85 turns out to work particularly well, and is the value used for PageRank computations in most cases.

Google was the first search engine to use PageRank, and many believe that most of its popularity is due to the success of this technique.

4.2 Link Spam Detection

With the increased use of link analysis and page popularity in search engines to improve query result precision, an increasingly large number of web-site designers have started developing techniques to mislead link-based ranking algorithms so to increase the rank of their pages undeservedly: these techniques are broadly known as *link spam*. This phenomenon is continuously increasing, and it is highly relevant from an economic viewpoint: Amit Singhal, principal scientist of Google Inc., estimated that the search engine spam industry might have had a revenue potential of \$4.5 billion in year 2004 if they had been able to completely fool all search engines on all commercially viable queries [3].

We might define a link-spam page as a page that is connected so to receive an undeservedly high rank by some search engine; this definition is, however, vague because it may depend on the specific query and on the specific search engine (or, more precisely, on the ranking techniques adopted by the search engine), and also requires a human evaluation about what "undeservedly high" means, something that may be difficult to define and even to decide. Nonetheless, as far as techniques like PageRank or similar static algorithms are concerned, link spam does not depend on the query and its focus is on trying to let a page appear more important or valuable than it is, thus deceiving the link-based algorithm.

As someone observed [3] a page might have an undeservedly high rank even if it is not a spam page; in a sense, the definition of link spam depends on the intention of the page designer, and as such it is relatively subjective: we may call spam the set of pages and links that a web-site designer would not have added to his/her site if search engines did not exist [22]. It is worth observing, as a side remark, that what we call "link spam" is also known as "search engine optimization" (SEO) by web designers, an expression that is aimed at inducing one to perceive spam as a good, or at least innocent, behaviour. We may also have good sites colluding each other to improve their ranking [2], so it is very difficult to distinguish ethical optimizations from bad ones. In particular, is it spam when good sites try to fight spam using techniques used by spammers?

As observed by many authors [2] [4], link spam is usually realized through the creation of a set of pages or domains with a highly-connected structure, often referred to as *link farms*, that are designed to convey rank towards some target pages; the links that are created with the only purpose of link spam are sometimes called *nepotistic*. Some authors, in particular, studied the impact of link farms on PageRank values [2] [8] [27], which turn out to depend both on the damping factor and on the structure of the link farm itself. A study of the optimal structure of link farms (a structure that maximizes rank gain) is presented in [13].

The latter paper is an attempt to give a taxonomy of link spam; as far as spammers are concerned, there are a number of domains over which they have full power, and where they can create hostnames and/or pages with almost no ex-

pense or effort. They also may have a limited access to other pages which they can moderately modify (e.g., blogs on which they can post comments). Finally there is a large part of the web over which they have no control whatsoever. Of course, spam detection is mostly based on the comparison between spam pages and the part of the web that is inaccessible to spammers. Some techniques used by spammers include creating a *honey pot* (a web site containing useful resources, e.g., manuals, that also contains links towards some target pages so as to boost their rank), infiltrating a publicly-accessible web directory, posting (usually, by means of a robot) comments in blog, message boards etc. that contain spam links, buying expired domains etc.

4.3 Detection and Penalization

While web content spam can be easily identified in most cases by human direct inspection, link spam is much more elusive [4]: pages designed to deceive ranking algorithms can look absolutely innocent and their content might appear to be valuable (often, it is indeed copy-and-pasted from existing non-spam pages). Trying to devise algorithms for the automatic *detection* of distrusted pages is absolutely necessary: some algorithms try to identify pages that are likely to be spam, i.e., that exist with the sole purpose of boosting the rank of some other pages, whereas other techniques focus on the search of nepotistic links, that is, links that mainly convey undeserved rank.

Once a link or a page has been marked as distrusted, we must find a way to *penalize* it so that its importance is reduced with respect to how it would be otherwise. The most radical solution is simply getting rid of the page, expunging it from the index forever and blacklisting it so that it will not be crawled in the future. The search engine can be more indulgent, and put the page in quarantine: the page will not appear in the index, but it will be crawled again in the future to give it a "second chance" of redemption. A more indulgent (but sensible and still effective) action consists in penalizing the page so that it receives a lower rank than it would with the ranking algorithm usually adopted.

Often, detection and penalization are really implemented as a single ranking technique that gives the same results as some other standard ranking algorithm (e.g., PageRank) for non-spam pages, while penalizing spam pages.

Of course, some simple yet sometimes effective solutions exist:

- Maintaining a blacklist of pages that are abusing the incoming links.
- Dropping links between pages with the same host, or even the same domain [8].
- Counting the number of host names under which a given IP address is known: when this number is very high, it is most likely a spam site [11].

Some by-now classic approaches to link-spam detection consist in trying to classify pages as spam/non-spam with the aim of some standard classifiers, using suitable page features, in the same way as e-mail spam is detected. One of the first works in this direction was [8], who tried to

characterize nepotistic links using the C4.5 classifier with purely syntactic features (such as whether the links were between pages with the same host/domain, the number of words in common between the titles of the two pages etc.), achieving a very high accuracy (above 90% in all cases). A similar approach was followed also in [1] in the more general framework of page categorization.

More recently Baeza-Yates et al. [2], in the same line of research, tried to combine a large number of link-based measures for spam detection, and used them as features for automatic classification via decision trees, obtaining an optimum spam-detection rate of about 79% with just 2.5% of false positives. The features they used include the page in- and out-degree, the reciprocity (fraction of out-neighbours of a page that are also in-neighbours), the assortativity (ratio between the out-degree of a page and the out-degree of its neighbours), the maximum PageRank of the site, the standard deviation of PageRank of the neighbours etc.

4.4 Diffusing Trustworthiness and Suspicion

Among the SEO community, it is believed [24] that some search engines (most notably, Google) are using a link-spam detection algorithm known as *BadRank*. This algorithm starts from a set of pages marked as "bad" and uses a technique similar to PageRank (but with arc direction reversed) to diffuse badness: the justification behind this algorithm is that a page is bad if it points to many bad pages (the PageRank slogan, instead, states "a page is good if it is pointed to by many good pages"). Even though it is common knowledge that this technique is adopted by Google, we don't know of any evidence behind this belief.

A conceptually similar, but dual, algorithm is TrustRank [14]: they start from a set of trusted nodes, and use a PageRank-like diffusion algorithm to spread the trustworthiness value all over the graph. Nodes with low trust value are then penalized as suspect.

A related idea [24] is to try to individuate link farms consisting of TKC (*Tightly Knit Communities*), following the same route as in [4] but obtaining an algorithm that can be applied fruitfully also to large (Web-size) graphs. They deem a page "bad" if and only if there are too many domains that happen to be both out- and in-neighbour of the page: the set of bad pages is called seed. Then, the seed is expanded recursively according to the idea that a page that points to too many bad pages is itself bad, as in BadRank. Finally, PageRank is computed by penalizing or deleting links from bad sites. We can also decide the quality of a page based on both measures: trust and spammicity.

4.5 Statistical Measures of Irregularity

A key observation that stands behind many spam-detection algorithms is that spam pages present some features that do not follow the usual statistical behaviour of non-spam pages; indeed this observation is the basis of the classifier-based methods described above. Some more recent observations [27] showed experimentally that there is a correlation between spamminess and the way in which

PageRank values react to changes in the damping factor: this remark suggests that one might compute PageRank for different values of the damping factor, and penalize the nodes that show a PageRank with suspicious behaviour. This approach has been often criticized in that it requires many re-computations of PageRank, and thus seems to be unfeasible, but in [5] a technique is presented that shows how to approximate PageRank for all values of the damping factor with a computational effort that is only slightly larger than that required for computing a single PageRank vector.

A very general algorithm based on the idea of statistical irregularity is SpamRank [3]. SpamRank acts in three phases:

- First of all, a Monte Carlo iterated algorithm is used to determine the set of supporters of each page (the supporters of a page P are those pages that give the highest contribution to P 's PageRank).

- Then, some measure of regularity is computed for the set of supporters of a given page; a simple measure of regularity is some statistical correlation (e.g., Pearson correlation) between the PageRank values and an ideal Zipf distribution: this measure is based on the empirical observation that PageRank values have such a distribution, and that this property is largely independent from the subset of the web graph under examination.

- Finally, if the set of supporters is very irregular, they are penalized proportionally to their irregularity.

5 Trends in AIR in the Web

The evolving nature of the Web leads to new killer applications and, in consequence, new forms of abuse. We want to show the challenges that Search Engine operators and Web users face in the next future, regarding these applications.

As more and more Social Network applications are getting more and more popular, they are also getting the attention of abusers. Some forms of marketing can be considered legitimate. For instance, Cinema Studios may use video sharing sites like YouTube for promoting their upcoming premières. This kind of marketing is legitimate, as users spread their interest as they usually do for other (amateur) contents.

But Social Network web sites are plagued by spam. There are specific forms of spam. For instance, as YouTube videos are presented to the user by using its middle frame, some abusers are uploading videos with advertisements in that frame. Or when people is informed that the Google Maps aeroplane is taking pictures of their neighbourhood, they try to get pseudo commercial messages recorded, as recently in Sidney (Australia). Other sites affected by spam are Wikipedia, the Open Directory Project, Flickr, Orkut, Digg, etc.

The very realization of Social Network Spam (recently named "*snam*") is the utilization of the FOAF (*Friend of a Friend*) message feature frequently found in this new genre of networks. Google's Orkut, a network of some 200,000

members, offers the ability to send messages to FOAFs. FOAF messages often contain conference promotions or job postings that, while low in volume, will one day require action on the part of network managers. In fact, this kind of spam is also affecting blogs.

The specialized forms of spam will require specific methods. However, we believe that current techniques used for addressing blog spam may well be useful for detecting and controlling Social Network spam, and no doubt they should be considered as the most promising starting point.

References

- [1] E. Amitay, D. Carmel, A. Darlow, R. Lempel, A. Soffer. The connectivity sonar: detecting site functionality by structural patterns. *HYPERTEXT'03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pp. 38–47, New York, NY, USA, ACM Press, 2003.
- [2] R. Baeza-Yates, C. Castillo, V. López. Pagerank increase under different collusion topologies. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.
- [3] A. Benczúr, K. Csalogány, T. Sarlós, M. Uher. Spamrank—fully automatic link spam detection work in progress. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.
- [4] K. Bharat, M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104–111, New York, NY, USA, ACM Press, 1998.
- [5] P. Boldi, M. Santini, S. Vigna. Pagerank as a function of the damping factor. *Proceedings of the 14th international conference on World Wide Web*, pages 557–566, New York, NY, USA, ACM Press, 2005..
- [6] S. Brunessaux, O. Isidoro, S. Kahl, G. Ferlias, A. Rotta Soares. NetProtect report on currently available COTS filtering tools. Technical report, *NetProtect Deliverable NETPROTECT:WP2:D2.2 to the European Commission*, 2001. Available at: <<http://www.netprotect.org>>.
- [7] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma. Adversarial classification. *Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Seattle, WA, USA, August 2004). KDD '04. ACM Press, New York, NY, 99-108.
- [8] B. Davison. Recognizing nepotistic links on the web. *Artificial Intelligence for Web Search*, pages 23–28. AAAI Press, 2000.
- [9] L. Denoyer, J.N. Vittaut, P. Gallinari, S. Brunessaux. Structured multimedia document classification. *DocEng '03: Proceedings of the 2003 ACM Symposium on Document Engineering*, ACM Press, 153–160.
- [10] R. Du, R. Safavi-Naini, W. Susilo. Web filtering using

- text classification. *Proceedings of the 11th IEEE International Conference on Networks*, 2003, Sydney, IEEE, 325–330.
- [11] D. Fetterly, M. Manasse, M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. *Proceedings of the 7th International Workshop on the Web and Databases*, pp. 1–6, New York, NY, USA, ACM Press, 2004.
- [12] J.M. Gómez, E. Puertas, M. Maña. Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization. *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data*, Palais du Grand Large, St-Malo / France, March 13-15, 2002.
- [13] Z. Gyöngyi., H. Garcia-Molina. Web spam taxonomy. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.
- [14] Z. Gyöngyi, H. Garcia-Molina, J. Pedersen. Combating web spam with TrustRank. *Proceedings of the 30th International Conference on Very Large Databases*, pp. 576–587, Morgan Kaufmann, 2004.
- [15] M. Hepple, N. Ireson, P. Allegrini, S. Marchi, J.M. Gómez. NLP-enhanced Content Filtering within the POESIA Project. *Fourth International conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, May 26-28, 2004.
- [16] P. Kolari, A. Java, T. Finin. Characterizing the Splogosphere. *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem, WWW Conference 2006*, <<http://www.blogpulse.com/www2006-workshop/>>.
- [17] P. Kolari, A. Java, T. Finin, T. Oates, A. Joshi. Detecting Spam Blogs: A Machine Learning Approach. *Proceedings of the Twenty-First National Conference on Artificial Intelligence*. July 16–20, 2006, Boston, Massachusetts. Published by The AAAI Press, Menlo Park, California.
- [18] P. Lee, S. Hui, A. Fong. A structural and content-based analysis for web filtering. *Internet Research* 13, 27–37, 2003.
- [19] G. Mishne, D. Carmel, D. Lempel. Blocking Blog Spam with Language Model Disagreement. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*.
- [20] A. Ntoulas, M. Najork, M. Manasse, D. Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006)*. WWW '06. ACM Press, New York, NY, 83-92.
- [21] L. Page, S. Brin, R. Motwani, T. Winograd. The PageRank citation ranking: Bringing order to the web. *Technical Report 66, Stanford University*, 1999. Available at <<http://dbpubs.stanford.edu/pub/1999-66>>.
- [22] A. Perkins. White paper: *The classification of search engine spam*, Septiembre 2001. Available at <<http://www.silverdisc.co.uk/articles/spam-classification/>>.
- [23] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1–47, 2002.
- [24] B. Wu, B. Davison. Identifying link farm spam pages. *Proceedings of the 14th International World Wide Web Conference*, Industrial Track, May 2005.
- [25] B. Wu, B. Davison. Cloaking and Redirection: A Preliminary Study. *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*.
- [26] B. Wu, B. D. Davison. Detecting semantic cloaking on the web. *Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May, 2006)*. WWW'06. ACM Press, New York, NY, 819-828.
- [27] H. Zhang, A. Goel, R. Govindan, K. Mason, B. Van Roy. Making eigenvector-based reputation systems robust to collusion. *Proceedings of the third Workshop on Web Graphs (WAW)*, volume 3243 of Lecture Notes in Computer Science, pages 92–104, Rome, Italy, Springer, 2004.

Abbreviations and Acronyms

- AIR: Adversarial Information Retrieval.
 ATC: Automated Text Categorization.
 FOAF: Friend of a Friend.
 HTML: HyperText Mark-up Language, a script used for writing the coding for web pages.
 ICRA: Internet Content Rating Association.
 IDF: Inverse Document Frequency.
 PICS: Platform for Internet Content Selection.
 POESIA: Public Open-source Environment for Safer Internet Access, <www.poesia-filter.org>.
 SEO: Search Engine Optimization, adjusting web-page scripts to achieve a high search engine ranking.
 TF: Term Frequency, number of times the token occurs in the page.
 TF.IDF: Term Frequency multiplied by Inverse Document Frequency.
 URL: Universal Resource Locator, a standard form of website address.