INFORMATION WASTE ON THE WORLD WIDE WEB AND COMBATING THE CLUTTER

Complete Research

Amrit, Chintan IEBIS Department, University of Twente, Enschede, NL, c.amrit@utwente.nl

Wijnhoven, Fons, IEBIS Department, University of Twente, Enschede, NL, fons.wijnhoven@utwente.nl

Beckers, David IEBIS Department, University of Twente, Enschede, NL, d.k.beckers@student.utwente.nl

Abstract

The Internet has become a critical part of the infrastructure supporting modern life. The high degree of openness and autonomy of information providers determines the access to a vast amount of information on the Internet. However, this makes the web vulnerable to inaccurate, misleading, or outdated information. The unnecessary and unusable content, which is referred to as "information waste," takes up hardware resources and clutters the web. In this paper, we examine the phenomenon of web information waste by developing a taxonomy of it and analyzing its causes and effects. We then explore possible solutions and propose a classification approach using quantitative metrics for information waste detection.

Keywords: Internet information waste, web spam, web site quality, Internet waste detection

1 .Introduction

The original purpose of the Web was to create a single, universal, accessible hypertext medium for sharing information (Berners-Lee & Fischetti, 1999). Over the years, the World Wide Web (WWW) has developed into a global information space with a multitude of autonomous information providers (Jacobs and Walsh, 2004). The different types and quality of WWW content have been a concern for a long time. The deficiency of enforceable standards has resulted in frequent information quality problems (Eppler and Muenzenmayer, 2002). A similar point is made by Arazy & Kopak (2011), p. 89): "With less traditional gatekeeping on the 'information production' side, more content is obtained from sources with mixed, and sometimes dubious, provenance." Not only the quality is a concern, the sheer amount of information is causing problems too. We appear to be in a state of "information overload" (Himma, 2007, Daradkeh et al., 2015). According to Bawden & Robinson (2008)(p. 183), "information overload occurs when information received becomes a hindrance rather than a help, even though the information is potentially useful." The amount of information on the WWW is increasing at an astonishing rate. With this information flood, the major task for information service providers has become one of filtering and selecting information rather than finding enough appropriate material (Bawden & Robinson, 2008).

A very prevalent form of information waste on the Internet is a phenomenon known as web spam. Araujo and Martinez-Romo (2010) (p. 1556) state that "Web spam, or spamdexing, includes all techniques used for the purpose of getting an undeservedly high rank." It has become common practice to craft pages for the sole purpose of increasing search engine rankings without improving the utility of the pages (Ntoulas et al., 2006). The most common web spam techniques include the artificial generation of content or keywords, cloaking, redirection spam, and link farms (Prieto et al., 2013). Broadly speaking, the following entities suffer economic losses from web spam (Prieto et al., 2013): (a) end users can be cheated and waste their time and money; (b) owners of web pages have difficulty in reaching their audience in an ethical way; and (c) search engine providers lose prestige and waste resources. Araujo and Martinez-Romo (2010) stated that web spam is one of the main problems of search engines because it strongly degrades the quality of search results. This leads to disappointment and frustration among users when they are finding spam sites instead of legitimate search results. With regard to the financial impact of web spam, spam site operators deprive legitimate sites of revenue that they could have earned via search engine referrals (Araujo and Martinez-Romo, 2010, Ntoulas et al., 2006). Moreover, search engine operators waste significant hardware resources (network bandwidth, storage space, CPU cycles) on crawling, processing, indexing, and matching to queries (Ntoulas et al., 2006, Prieto et al., 2013). It must also be taken into consideration that combating web spam requires substantial investment in manpower and time to keep search engines usable.

Information waste is closely related to poor *information quality* (Wang and Strong, 1996). Low-quality information could be considered information waste because it will be unusable and could be considered unnecessary. Information quality has been commonly defined as the fitness for use of information (Bizer and Cyganiak, 2009). Taylor (1986) identifies five attributes of information quality: accuracy, comprehensiveness, currency, reliability, and validity. Knight and Burn (2005) collate 12 widely accepted information quality frameworks from information systems research containing 20 information quality attributes. A crucial point however, that has been mentioned is that information quality needs to be assessed within the context of its generation and use because the attributes of quality can vary depending on the context in which it is to be used (Knight & Burn, 2005). "What is of good quality in one situation might not be of much help in another situation" (Mai, 2013). Quality dimensions such as relevance and usefulness are thus difficult to measure reliably due to their subjectivity (Knight & Burn, 2005). In support of this point, Arazy & Kopak (2011), who did empirical research on rating information quality attributes, found that it was difficult to reach agreement on the assessment of quality dimensions. Arazy & Kopak conclude that information quality (2011, p. 98): "… is an elusive construct that is hard to measure, and users' quality estimates are subjective, therefore making it difficult for multiple assessors to reach an agreement on a resource's quality." Similarly, Mai (2013) stated that information quality is a difficult concept to quantify effectively, even though significant research effort has gone into developing objective characteristics.

Information trust is also related to information waste. Trust is the intervening variable that affects the use of technology, mediating between information quality and usage (Kelton et al., 2008). Because there is a lack of standard procedures and editorial controls, it is difficult to create trust in online environments (Kelton et al., 2008). As the internet becomes more pervasive in our lives, trust in digital information becomes more important because it plays a key role in some of our decision-making processes and personal knowledge. There might even be harm in relying on poor information.

Hence, dealing with *information waste* on the Internet is an important issue. We use the following definition of information waste: "*information which is unnecessary (e.g. redundant) and unusable (e.g. not understandable) and which are the consequences of human limitations of knowing which data are of no use and could thus be removed or stored on a non-direct access medium*" (F Wijnhoven, Dietz, & Amrit, 2012, p. 135). Information waste in the context of the World Wide Web is pertinent, given the multitude of quality concerns and the rapidly expanding amount of information. The information flood challenges internet users with an abundance and redundancy of web resources, which makes it difficult to optimize search, identify dependable sources, or obtain factual information from the clutter (Langford, 2010). Furthermore, it might be harmful to rely on poor information, as this may have an impact on decision-making, personal knowledge, and reference materials (Kelton et al., 2008). Being able to filter out or eliminate a significant portion of information waste will reduce "information overload" as well as provide more reliable and trustworthy information. Moreover, less hardware resources will need to be used, which has environmental benefits and potentially reduces the operational costs of the internet (Wijnhoven, 2012, Wijnhoven et al., 2014). In this paper we pose two key questions: (1) What variants of information waste exist? (2) Is it possible to determine the potential information waste on the Internet? In an attempt to answer question 2, we propose an approach for detecting Internet waste employing objective user data. Our task in this paper is essentially to predict the value of websites. Hence, our task is primarily one of prediction rather than one of explanation (Shmueli, 2010). Therefore we perform a predictive analysis (Shmueli and Koppius, 2011) rather than a statistical explanatory analysis (as attempted above) to create a predictive model (Gregor, 2006).

2. A taxonomy of information waste

To understand the diversity of information waste, we identify it at technical and human levels of the meaning of the term information, following a semiotic perspective. Semiotics (from the Greek word for sign) is the doctrine and science of signs and their use (Brier, 2005). A number of scholars have suggested establishing the foundations of information studies in semiotics (Mai, 2013). Semiotics allows for a more nuanced view on technical and human levels of information concepts. Stamper's (1991, 1996) semiotic ladder provides an important means for understanding both the physical and social dimensions of information. Information has a distinct meaning that can be assigned to it at each ladder of signs from technical to human use. The semiotic framework consists of the following six layers (Stamper, 1991): (1) the physical world; (2) empirics; (3) syntactic; (4) semantics; (5) pragmatics; and (6) the social world. Boell and Cecez-Kecmanovic (2010) assume a continuum from the physical/empirics to the pragmatics/social layers, which is slightly different from Stamper's (1991) framework. Attributes closer to the physical world are more closely associated with technical solutions while attributes closer to the social world focus on information use and how they influence users' actions.

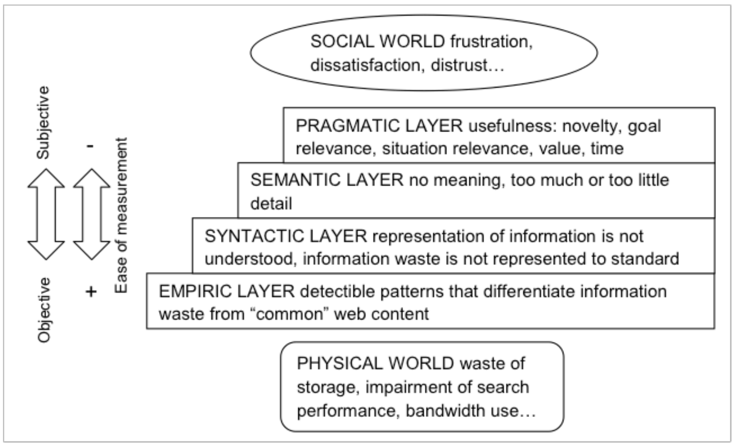
SOCIAL WORLD frustration, dissatisfaction, distrust…
PRAGMATIC LAYER usefulness: novelty, goal relevance, situation relevance, value, time
SEMANTIC LAYER no meaning, too much or too little detail
SYNTACTIC LAYER representation of information is not understood, information waste is not represented to standard
EMPIRIC LAYER detectible patterns that differentiate information waste from "common" web content
PHYSICAL WORLD waste of storage, impairment of search performance, bandwidth use…
Subjective
Objective
-
+
Ease of measurement

Figure 1: The semiotic framework of information waste

The semiotic framework of information waste (Fig. 1) is an extended and adapted version of the semiotic framework proposed by Stamper (1991) and the subsequent reinterpretation by Boell & Cecez-Kecmanovic (2010). The models have been adapted to get a comprehensive overview of *information waste* according to the level of structure we give to signs. The physical world layer and the social world layer are shown as separate variables. The semiotic layers form steps between these two worlds and information waste is given a precise meaning according to the level of structure. Each layer contains a short description. The two arrows on the left-hand side of the model indicate the nature and associated challenges of analysing web content at each semiotic level. Measures for determining the value and quality of a web site are *objective* in the lower semiotic layers, which allows programmatic evaluation by applying established heuristics. There is a relatively high *ease of measurement* in these lower layers. When moving up to higher layers, measurements are harder to evaluate automatically, thus requiring human interpretation. In the pragmatic layer, judgments are highly dependent on personal preferences, making evaluation highly subjective. In addition, obtaining reliable computerized proxies for information quality is difficult. In the following subsections we present a literature search within the areas of web spam and related terms to provide a selection of information waste detection methods for each of the semiotic layers. We discuss concepts like "goal relevance" and "trust" as well as the operational definition of information waste, when they are relevant to the particular semiotic layer being discussed.

2.1 Physical world layer

The physical world is the layer consisting of the physical phenomena shaping our everyday lives. Information waste affects the physical world because information needs a physical carrier and thus has physical effects (Schmidt et al., 2009). These effects are mainly the unneeded use of hardware resources and extra effort needed to sort and filter information. Unneeded and unusable web content takes up storage space that could have been used for better purposes. Also, bandwidth is wasted when bad content is (often inadvertently) accessed. The second physical effect of information waste is that a lot of effort is needed to detect bad content and maintain search engine performance (Wijnhoven et al., 2012). Algorithms have to be updated constantly to keep up with the stream of information waste. Heavy investment is also needed in hardware resources to cope with these demands.

2.2 Empiric layer

To inform, information needs to be detectable. This requires it to be distinguishable from background noise. Information is distinguishable from background noise when patterns can be detected. If there are no distinguishable patterns, there is no message, and there will not be information. Information waste, especially web spam, follows specific patterns that differentiate it from legitimate web content. The general patterns used for web spam detection observe links, the rate of evolution of pages, and behavioural patterns. Detecting questionable websites on the basis of empiric cues has been widely researched as it can be automated and leads to consistent results.

Operational definition of information waste for the Empiric layer: information waste follows detectable empirical patterns that differentiate it from "common" web content.

One common method for detecting questionable web sites is by using link patterns. Web graphs provide a structural signature of sites (Diligenti et al., 2000); a high density of connections is associated with link farms (Geng et al., 2007). Moreover, having far more external links than internal links, many broken links, and the presence of many common pages in links are found to be useful spam indicators (Araujo and Martinez-Romo, 2010, Geng et al., 2007). Evolution patterns also tend to differ between legitimate and questionable sites. Shen et al. (2006) therefore proposed that drastic changes in the number of links leading to a site as an indicator of web spam.

2.3 Syntactic layer

Syntactics observes how signs relate to other signs (Ryder, 2005), denoting the representation of information. Information needs to be represented in a certain form – a set of principles and rules. In other words, it needs to be represented using a syntax understood by the recipient. This is not limited to the syntax of natural language; it can also be the layout of a page or the scripts running on it. Syntactic attributes are widely used to detect information waste. Information waste is represented in ways that differ from the way legitimate content is displayed. The main categories for detection methods of web spam in the syntactic layer are language-based indicators, source code features (including layout), and content features. Regarding content, it is worth noting that in the syntactic layer only the representational components of the content are examined, not the meaning.

Operational definition of information waste for the Syntactic layer 1: Information waste is information represented in a way that makes it incomprehensible to recipients.

Operational definition of information waste for the Syntactic layer 2: Information waste is information represented in a way that differs from the way legitimate information is represented.

Some research has looked into code features of web sites for spam detection. Some programmatic features are commonly used to produce web spam and some code functions also hide it to make it harder for search engines and site administrators to find. For example, HTML injection and cross-site scripting lead to recognizable features in the coding of a website, and programmers often try to hide redirections, function or content by codifying this in a certain manner (Prieto et al., 2013). In addition, spam site attributes such as the size in bytes, number of words, and the relation of code to site content all tend to differ from normal web sites. Some content features are also syntactic, and have been used for spam detection as well. Some pages contain text that had been generated by spinning other articles. This results in odd grammar and vocabulary from popular words to attract search results. This can be detected by using *n*-grams and certain text typologies (Araujo and Martinez-Romo, 2010, Ntoulas et al., 2006). Empty chains or links which only contain punctuation marks or numbers have also been proposed for spam detection because they are clearly not intended to be used normally (Araujo and Martinez-Romo, 2010).

2.4 Semantic layer

"Semantics studies the affiliations between the world of signs and the world of things" (Ryder, 2005). A message needs to be comprehensible to the recipient for it to be meaningful. The message has to be integrated into a recipient's knowledge for it to become information. If it cannot be integrated, it will not be understood. Too little specificity and depth will not sufficiently inform the recipient, so the message will not be fully understood. On the contrary, too much specificity and depth overwhelms the recipient and the message will also not be understood. Automatic evaluation of semantic attributes of web pages is only possible to a limited degree. There are some methods that can measure the coherence of a web page as well as other language-based features. Algorithms can only make some inferences about the content that can partially distinguish an informative web page from a page created for malicious intent.

Operational definition of information waste for the Semantic layer 1: Information waste is information that has no meaning to the recipient.

Operational definition of information waste for the Semantic layer 2: Information waste is information that has too much or too little detail to serve the user.

Operational definition of information waste for the Semantic layer 3: Information waste is information that aims to mislead or spam the user.

One method of detecting web spam semantically is by assessing the thematic nearness. A large divergence between a link's anchor text and the text of the linked page or disagreement of the title and content of the page can be a spam cue (Sharapov and Sharapova, 2011, Araujo and Martinez-Romo, 2010). Some methods have also looked at the content of web spam pages from a semantic perspective for example by querying certain words and phrases associated with spam. It is also possible to measure the degree of coherence and grammatical and semantic sense of web pages (Prieto et al., 2013, Sharapov and Sharapova, 2011, Wang et al., 2010).

Wang et al. (2010) developed a content-based trust model for spam detection. By adding information quality measures to their normal web spam detection method, they managed to significantly increase the reliability of their method. The information quality measures included currency (last modification time), information-to-noise ratio, and popularity. These variables serve as proxies for pragmatic metrics and are suitable for automatic analysis.

2.5 Pragmatic layer

Pragmatics explains the effect of signs on human behaviour. These larger structures have a purpose in human communication. Information at this level leads to intentions and actions. For this, the information needs to be useful and valuable to the user. The pragmatic layer therefore consists of several attributes.

Operational definition of information waste for the Pragmatic layer 1: Information waste is not useful and not valuable to the recipient.

Novelty character of information: informing a recipient of something new is a central attribute of information. A message informing someone of something he or she already knows does not make the recipient any more informed. This is redundancy. However, novelty is not always essential, and redundant messages can serve as a helpful confirmation in some cases.

Operational definition of information waste for the Pragmatic layer 1.1: A message is information waste when it does not provide novel information and does not serve as a needed confirmation.

Goal relevance: information must help its recipients to make informed decisions by making sense of situations. Information that can be used to achieve a goal or make an informed decision has relevance to its users. Achieving goals and making informed decisions implies that the information required must be sufficiently accurate and complete for the task at hand.

Operational definition of information waste for the Pragmatic layer 1.2: Data that does not help its recipients make informed decisions is information waste.

Situational relevance: information may only be useful in certain situations. For example, the gas prices at nearby gas stations will only be useful when you need to refuel your car.

Operational definition of information waste for the Pragmatic layer 1.3: Information waste is information that is not relevant to the current situation.

Trust: information needs to be trusted by the recipient before he or she takes any actions that depend on it. "Perceived trustworthiness of information can be evaluated in terms of its accuracy, objectivity, validity, and stability" (Kelton et al., 2008). Trust can relate to both the content itself and the information source. The qualities of the information source are referred to as credibility (Savolainen, 2011). If someone perceives information not to be trustworthy, decisions and actions will be delayed.

Operational definition of information waste for the Pragmatic layer 1.4: If information cannot be trusted, it is information waste.

Value to a recipient: value of information can be narrowly defined as instructional and economic value. Instructional value helps people or organizations to make decisions or solve problems. Economic value allows an individual or organization to make profit or avoid costs. *Operational definition of information waste for the Pragmatic layer 1.5:* Information is waste when the recipient is unable to obtain any instructional or economic value from it.

Time dependence: Something might only be information at a certain point in time, while being irrelevant at another time. For example, knowing when the next bus home will leave is very relevant at the end of the day when it is time to go home after work. This information is less useful when you do not intend to go home yet.

Operational definition of information waste for the Pragmatic layer 1.6: A message provided outside the time during which it is required is information waste.

Developing methods for the detection of pragmatic information waste is difficult because of the subjective and time and context dependence of it. The literature search only found one research paper that attempts to examine pragmatic aspects of information waste. It is also notable that they used proxies for pragmatic attributes rather than truly pragmatic indicators.

2.6 Social world layer

At the highest level of the model, there is the social world. This layer consists of the information structures that constitute our social existence. This layer is affected in a large part by the communications we have with other human beings. Information waste has the effect that people have more difficulty in finding what they are looking for, and online information providers might not be trusted. Web spam and related phenomena have been plaguing the Internet for a long time. The extra effort needed to handle bad content is a form of waste as well, in line with Hicks' (2007) perspective on information production within a corporate setting. Moreover, false and misleading information has a negative impact on online transactions (e.g. e-commerce) as it reduces trust. Spammers have also misused the recent proliferation of web 2.0 platforms, and information quality problems have been more rampant than ever before.

3 Internet waste detector proof-of-concept method

3.1 Kernel idea

Existing Internet waste detection methods intend to find web spam and fraudulent e-commerce sites, which are classified as information waste due to their empiric and syntactic properties. Efforts to do something about semantic, pragmatic and social information waste have been limited so far. One of the reasons for this is that it is difficult to reach consensus on the true meaning of important concepts such as “relevance” and “information quality” (Hjørland, 2010, Mai, 2013). Relevance and information quality were found to be highly subjective and therefore difficult to evaluate automatically. Solid theories of objective and subjective relevance are needed, yet they are complex and riddled with paradoxes (Hjorland, 2010).

Wijnhoven & Amrit (2010, 2014) propose a subjective file questionnaire to determine the semantic, pragmatic and social value of files in a file system. However, filling in such a questionnaire is too labour-intensive, so a method for automatically identifying file value is proposed. Five empiric and syntactic file characteristics were proposed as determinants of file value: (1) frequency of access; (2) file age; (3) last modification time; (3) file type; and (4) user grade (rank of person using the file). They suggest that if correlations between these propositions are corroborated, the file characteristics can be used as decision parameters in a file retention method, while rejected propositions should be excluded from the data retention policy.

We use this basic theory of predicting the objective value using subjective data for Internet content. Web analytics are a popular method to gather empiric and syntactic data in order to improve the effectiveness of web sites (Kent et al., 2011). Web analytics are gathered with the specific purpose of optimizing web content in order to make it more (pragmatically) valuable and (semantically) useful to users. For example, data gathered by Google Analytics can help web site owners determine which pages are the most popular, what type of information the users are looking for, and how much time they spend on the site (Google.com, 2014). It can therefore be assumed that web analytics are indicative of pragmatic attributes. A major advantage of web analytics is that the information is relatively more objective and can easily be obtained. We propose access speed (Eppler and Muenzenmayer, 2002, Palmer, 2002, Yang et al., 2005), the number of incoming links (Palmer, 2002, Yang et al., 2005), frequency of access, time on site, bounce percentage, global page view percentage, and global user percentage as objective site characteristics. Palmer (2002) developed and validated subjective characteristics: (1) amount of information; (2) ease of obtaining information; (3) information from other pages; (4) usability of the website; (5) layout quality of the site; (6) speed of site; (7) quality of information display; and (8) if people see it as valuable to return to the site. These characteristics resemble the attributes of the pragmatic layer, but they can be assessed more consistently as they apply specifically to web sites.

It is not yet clear to what extent these objective (syntactic and empiric) measures are indicative of semantic and pragmatic subjective value. If a high predictive relationship between subjective assessment and objective indicators can be found, these objective indicators can be used as proxies for subjective value and information waste. Figure 2 shows the predictive model to be used for this study. The objective, empiric metrics are gathered via web analytics tool that tracks the usage behaviour of web users. The subjective value is comprised of the attributes of the pragmatic layer of the semiotic framework. We expect that the way Internet users use web pages is indicative of the subjective value they attach to pages. Using objective attributes, a classifier will eventually be given the task to identify sites scoring low on subjective attributes. Depending on their effectiveness, objective measures can eventually be used for the development of a tool capable of automatically determining whether certain web content is waste.

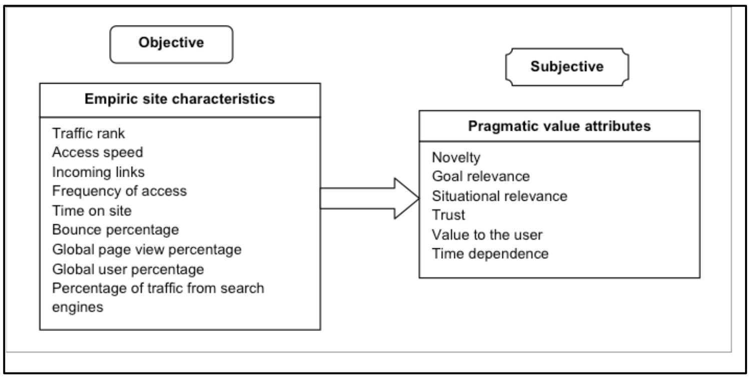

Figure 2: Our model for predicting pragmatic value attributes from empiric site characteristics

3.2 Dataset

3.2.1 Objective Metrics

Objective analytics from websites can be obtained in various ways. Site analysers are scripts installed on web servers to provide web administrators with objective site performance metrics. An alternative is to crawl the Internet and generate metrics from the crawled data. Another option is to install a browser add-on to track users' behaviour. Each of these methods has certain disadvantages for analysing information waste on the web. Using site analysers requires that every server has the same script if a representative portion of the web needs to be examined. Crawling a large portion of the web requires advanced hardware and is time-consuming. The third option, a browser add-on, may not provide representative data, as only a limited number of people might install it. Internet users also generally object to being tracked extensively.

Because of these issues in generating objective metrics by ourselves, we use the data provided by Alexa.com. Alexa.com provides analytic resources to web developers and administrators. Alexa.com gives an assessment of a web page by collecting the access speed, the number of incoming links, frequency of access, time on site, bounce percentage, global page view percentage, and global user percentage. The data is mainly gathered via a toolbar that tracks user behaviour. Alternatively, web site administrators can install site analyser scripts on their servers to gather data. Alexa.com does not provide an entirely representative overview of the Internet, as their metrics are only accurate for the first 100,000 web pages in their ranking. Beyond this list, there are not enough site visitors to provide accurate statistics. Nevertheless, the sites considered in Alexa's ranking are approximate to what users will typically perceive. Personal preferences and search engine referrals make sure that only a user sees only a certain portion of the web. The following metrics are available via Alexa.com (Fons Wijnhoven, 2012):

- *Traffic rank:* this rank is based on the traffic data provided by the toolbar panel over a rolling three-month period. A site's traffic rank is composed of a combined measure of unique visitors and page views ("How are Alexa's traffic rankings determined? – Alexa Support," n.d.)[1].
- *Access Speed* gives an indication of whether a site feels "fast." If a site feels slow it is more likely that users might leave the site. Loading speed is also indicative of the technical condition of a web site.

[1] https://alexa.zendesk.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined-

- *Links.* If a website has a lot of incoming links, it is expected to contain good information. When other sites link to the site, it could be because the site contains useful or valuable information. Linking can be seen as a form of endorsement, and this principle is one of the foundations of Google's PageRank algorithm (Brin and Page, 1998).
- *Frequency of Access* is the number of unique monthly visitors to a site. If a site has a lot of visitors it most likely contains valuable information.
- *Time on site*. If a user stays at a site for a long time, it most likely is good information because the user takes time to read the entire page. Precautions need to be taken with this metric because if a user keeps his browser open at a certain page while he is away or browsing other sites, it will give a false positive. Users also spend a long time on sites such as email and social networks to receive notifications while doing other things.
- *Bounce percentage* is the percentage of unique users who visited only one page on a certain website. If a user only visits one page, it could mean that the page (and the rest of the website) are not interesting. However, it could also be that the page is not exactly what he/she was looking for; the information quality might not necessarily be bad.
- *Global page view percentage*, which gives the percentage of pages viewed from a website compared to the estimated total number of page views.
- *Global user percentage*, which gives an estimation of the percentage of global internet users who visit a specific site.
- *From search engine* indicates the percentage of users who visited the site via a search engine. If a higher portion of users comes by entering the URL directly, it could mean that they frequently use the site, therefore being of significant value to them. On the other hand, if many people find the site via a search engine, it could mean that the website has a good search engine ranking and seems relevant to many users.

These metrics are not always conclusive for the assessment of web pages; therefore, a combination of these indicators needs to be used. Only sites in English were selected because a user survey was needed. To ensure that we select a random sample, the top million sites listed by Quantcast.com were used to randomize the sample. One hundred sites were selected from this pool by dividing the complete set into 100 equal subsets, from which one site was selected each time.

3.2.2 Subjective metrics

To obtain subjective ratings, we developed a feedback tool. The tool allows someone to open a web page and rate it on a scale of 1 to 5 along the dimensions of content, relevance, and comprehensiveness. The process described in the end of the previous section resulted in a sample of 150 sites selected randomly. One researcher rated all 150 sites. The first two authors then also independently rated a smaller subset of thirty sites and the inter-rater reliability was computed for verification. Kappa inter-rater reliability tests (Cohen, 1960) are typically performed between two individuals; therefore, we computed two kappa values. We then used Landis and Koch's (1977) proposal for interpreting kappa values to assess the values we obtained. Cohen's kappa was 0.24 in one case and 0.22 in the other case, both indicating fair agreement according to Landis and Koch's interpretation proposal.

3.2.3 Classifier algorithms

Predicting information waste with our model (Fig. 2) is a classification problem – a specific object is placed in a set of categories, based on the respective object properties (Gorunescu, 2011). The objective of classification is to analyse historical data stored in a database and automatically generate a prediction model that can predict future behaviour (Turban et al., 2011). In the first stage of the classification process, a classification model is constructed by applying an algorithm on the training set. In this classification model development stage, the chosen model adjusts its parameters starting from the correspondence between input data and corresponding known output (Gorunescu, 2011). The induced model consists of generalizations over the records of a training dataset, which help distinguish predefined classes (Turban et al., 2011). Once the classification function is identified, the accuracy can be verified using the testing set by comparing the predicted output with the observed output. Classification models are typically compared to other models and algorithms to find the best one for the situation. There are four categories of classification algorithms in general: naïve Bayesian, clustering, decision trees, and neural network classifiers (Han et al., 2006). Ensemble learning techniques such as boosting and Random Forests combine multiple classifiers to increase the accuracy of classification.

Three classification methods will be considered in this paper: Classification and Regression Trees (CART), Support Vector Machines (SVMs), and Random Forest. Each of these classification algorithms has different strengths and weaknesses and it is therefore worthwhile to compare their performance. *CART* has the main advantage that it is relatively robust to outliers and noise. CART is also quite intuitive because the models can be visualized and the underlying principle is not excessively complex. However, the weakness of CART is that its structure can be unstable; slight changes in the training set can lead to dramatic changes in the decision tree (Kantardzic, 2011). The principle behind *SVMs* is based on the solid theoretical background of statistical learning which can effectively handle statistical estimation with small samples (Kordon, 2009). SVMs are currently one of the fastest-growing approaches of computational intelligence. SVMs create a hyperplane that splits the data into two parts. The support vectors are the vectors that lie on the margin of the hyperplane. These vectors are then used to define the decision rule or model. Some of the main advantages of SVMs are explicit model complexity control, repeatable results, and solid theory. The disadvantages of SVMs are that the approach is extremely mathematical and complex. Furthermore, the experience of SVMs in large-scale industrial applications and model support is relatively limited (Kordon, 2009). *Random Forests* are an example of an ensemble learning method. Random Forests were introduced by Breiman (2001) and serve as an extension of his bagging idea and were also developed as a competitor to boosting (Cutler et al., 2012). Random Forests combine the results of various predictive models generated during training (Kantardzic, 2011). Correct decisions are reinforced when there are multiple independent "decision-makers." Ensemble learning is a promising approach for improving the accuracy of a predictive model.

3.2.4 Classifier training

We used the following objective metrics in our classification model, as previously described in section 3.2.1: (1) Traffic Rank; (2) access speed; (3) number of incoming links; (4) frequency of access; (5) time on site; (6) bounce percentage; (7) global page view percentage; (8) global user percentage; and (9) percentage of users from search engines. The initial dataset was pre-processed to give the classifier as little confusion as possible. Access speed was missing in many instances, so this attribute was removed altogether. This was judged to be appropriate because access speed relates more to the "feel" of a website rather than information quality. Subjective measures for each page pertain to the content, relevance, and comprehensiveness and were expressed with a number from 1 (low) to 5 (high). These ratings were added up to form a composite measure. We considered web sites that fell below a certain rating (threshold will be discussed later) as information waste, while web sites with a rating above this threshold were labelled as non-waste. Our dataset consists of 150 complete records, which is not very large, yet sufficient to train a classifier (Alpaydin, 2004).

A training set with an equal number of waste and non-waste instances will create a classification model that is not biased towards the class that appears more often. We trained the classifiers by using 10-fold cross-validation, also known as rotation estimation. This is more advanced that the simple split methodology and more suitable for a small training set. Positive and negative precision and positive and negative recall were used as metrics for performance evaluation, resulting in a four-way classification of the results:

- *True positive:* a waste page classified as waste
- *True negative:* a non-waste page classified as non-waste

- *False positive:* a non-waste page classified as waste
- *False negative:* a waste page classified as non- waste

Given this four-way classification of results, the metrics are calculated as follows:

- *Positive precision* is the number of true positives as a fraction of all the waste classifications
- *Negative precision* is the number of true negatives as a fraction of all the non-waste classifications
- *Positive recall* (or *Sensitivity*) is the number of true positives as a fraction of all the true waste pages
- *Negative recall* (or *Specificity*) is the number of true negatives as a fraction of all the true non-waste pages

For thresholds at 0.25 intervals between 2.5 and 5.5, we created a separate dataset consisting of: the class label (according to the threshold) and, the objective measures. Using these sets, we trained each of the classifiers: CART decision tree, SVM, and random forests. The training was done by randomly taking a 47% sample (70 of the 150 websites) of the data. Next, we used the trained classifier to test the remaining 53% (80 of the 150 websites). We then compared the results of the classification to the actual labels.

3.3 Results

We used R (Bunn and Korpela, 2012) and its related packages to generate classification models from the training set. Using all Alexa variables on a training set consisting of 80 records, the classifiers perform as shown in Table 1.

Figure 3 shows the performance of the CART classifier in terms of precision, recall, and accuracy as a function of threshold. This shows that the threshold that maximizes the precision, recall, and accuracy is around 4.7. We get similar plots for Random Forests and SVM.

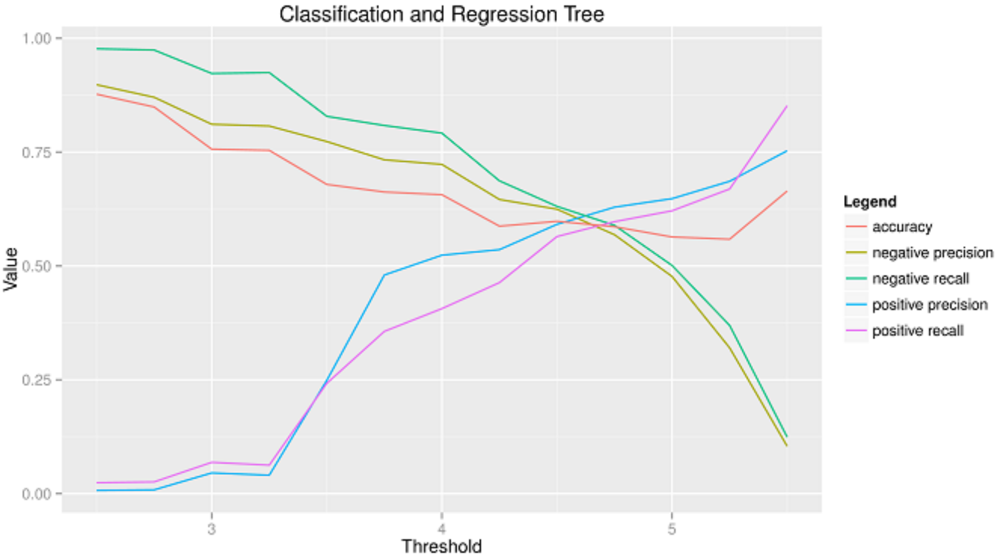

Figure 3: Finding the optimum threshold for waste indication

To illustrate the use of the classifier, we developed a tool that uses the Alexa "top million" list and classified all the websites on this list. To enable larger scale data gathering, we built a tool with a simple user interface that shows the user which website is being processed, what the current ratio is, and how many websites have already been processed. From our initial sample of 70 web sites that had objective values, and a waste threshold of 4.75, around 43 were still considered potential information waste giving an overall potential information waste rate of 61% (Table 1) for the Random Forest classifier.

Classifier	Negative Precision	Positive Precision	Negative Recall (*Specificity*)	Positive Recall (*Sensitivity*)	Accuracy
Random Forest	0.57	0.61	0.55	0.62	0.58
CART	0.57	0.63	0.59	0.60	0.59
SVM	0.59	0.68	0.66	0.58	0.59

Table 1: Classifier training results at a waste threshold of 4.75

3.4 Discussion

Table 1shows that the accuracy of the Random forest classifier is 58%, while it is slightly higher for CART and SVM classifiers (59%). Since this is just a proof of concept, such a result can be considered a good first step.

This paper provided a number of subjective metrics able to measure the information value of a website. Furthermore, we provided a number of objective metrics, which could easily be scraped from a website.

Using these metrics the answer to our question of *whether it is possible to develop a tool that can identify potential information waste on the Internet, is yes.* This paper provided a proof-of-concept to able to identify potential information waste on websites with an accuracy of about 59%.

We developed a tool that checked 500,000 sites (half a million), consisting of a random sample of Alexa.com's top 1 million sites that contained non-null data. We now ran the Random Forest classifier over the sample of half a million pages and found 52% was classified as potential waste (which is also the probability that a web page is classified as waste: *P(WebPage=Waste)*). If we now correct for classification error and use the formula suggested by Hopkins and King (2010) (page 235):

$$P_{actual}(WebPage = Waste) = \frac{P(WebPage = Waste) - (1 - Specificity)}{Sensitivity - (1 - Specificity)}$$

where sensitivity and specificity are defined earlier in 3.2.4, and using the values of Sensitivity and Specificity from Table 1, we get the corrected percentage of waste websites as 41% in the sample. Since these websites were only from Alexa.com's top 1 million websites, future research could expand the sample space. In order to develop an improved information waste detector, a larger sample set should be created. This would allow better training for the classifier. The objective metrics do have their limitations. The metrics were selected based upon what the scraper could provide. Further research could resolve this issue, by providing other metrics.

The classification model can also be extended with other variables. The current model was limited due to its objective metrics source. A good alternative might be Urlspy.co.uk that provides among other features, the number of pages on a website and the amount of external links (outgoing).

4 Conclusion

The openness and freedom to provide information of all kinds on the World Wide Web has led to frequent information waste problems. The web is riddled with false and irrelevant information, and this is increasingly putting a burden on Internet users. Our literature review has also shown types and causes of information waste in order to get an understanding of the extent of the problem. We have used semiotics for creating a taxonomy of information waste. Semiotics can be said to form the foundation of information studies. This helps to categorize and develop relevant information waste detection methods, which may hold the key to a World Wide Web that is less cluttered with irrelevant and inaccurate information.

There are very few examples of methods that attempt to detect the type of information waste described in this paper. Existing literature has mainly focused on detecting web spam, fake sites, and fraudulent e-commerce sites. In this paper, we propose a novel approach for detecting information waste using web analytics through the development of a predictive model (Gregor, 2006, Shmueli and Koppius, 2011). The basic reasoning is that the way users use a web page is indicative of its value and usefulness, which are pragmatic and semantic within the semiotic framework. We expected a link (Fig. 2) between objective empiric and syntactic metrics (produced by web analytics) and semantic and pragmatic subjective metrics (produced by subjective rating of sites). This link was partially established (Table 1) by applying data mining techniques for classification. Though the performance of our predictive model is low for reliable implementation as a waste detector, we can estimate the total number of waste webpages by correcting for classifier error (Hopkins and King, 2010). However, more objective variables and a larger dataset are needed to improve the classifiers. Additional data like the click through data for the websites as well as ways to handle noise in this data (Van Der Spoel et al., 2013) can help in improving the classification accuracy.

A well and widely used information waste detector could give website owners the proper kind of feedback for maintenance efforts or removal of content, both likely resulting in a much higher appreciation of Internet content or an increased efficiency of search and resource utilization. The actual impact and usefulness and social value of a well working information waste detector of still needs further research.

References

ALPAYDIN, E. 2004. *Introduction to machine learning*, MIT press.

ARAUJO, L. & MARTINEZ-ROMO, J. 2010. Web spam detection: new classification features based on qualified link analysis and language models. *Information Forensics and Security, IEEE Transactions on,* 5**,** 581-590.

ARAZY, O. & KOPAK, R. 2011. On the Measurability of Information Quality. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,* 62**,** 89-99.

BAWDEN, D. & ROBINSON, L. 2008. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science,* 35**,** 180-191.

BIZER, C. & CYGANIAK, R. 2009. Quality-driven information filtering using the WIQA policy framework. *JOURNAL OF WEB SEMANTICS,* 7**,** 1-10.

BOELL, S. & CECEZ-KECMANOVIC, D. 2010. Attributes of information.

BREIMAN, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

BRIER, S. 2005. *Semiotics: Nature and Machine* [Online].

BRIN, S. & PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems,* 30**,** 107-117.

BUNN, A. & KORPELA, M. 2012. R: A language and environment for statistical computing.

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement,* 20**,** 37-46.

CUTLER, A., CUTLER, D. R. & STEVENS, J. R. 2012. *Random Forests,* Berlin, Heidelberg, Springer.

DARADKEH, Y., SELIMI, E. & GOUVEIA, L. 2015. Information Overload: How to solve the problem? Current trends in technology and its impacts to individuals and organizational context. *International Journal of Open Information Technologies,* 3**,** 27-30.

DILIGENTI, M., COETZEE, F., LAWRENCE, S., GILES, C. L., GORI, M. & OTHERS 2000. Focused Crawling Using Context Graphs. *VLDB.*

EPPLER, M. J. & MUENZENMAYER, P. 2002. Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology. *IQ.* MIT.

GENG, G.-G., WANG, C.-H., LI, Q.-D. & ZHU, Y.-P. 2007. Fighting link spam with a two-stage ranking strategy. *Advances in Information Retrieval.* HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY: SPRINGER-VERLAG BERLIN.

GOOGLE.COM. 2014. *Google Analytics Official Website – Web Analytics & Reporting* [Online]. [Accessed 21-11-2014 2014].

GORUNESCU, F. 2011. *Data Mining: Concepts, Models and Techniques,* Berlin, Heidelberg, Springer Berlin Heidelberg.

GREGOR, S. 2006. The nature of theory in information systems. *MIS quarterly***,** 611-642.

HAN, J., KAMBER, M. & PEI, J. 2006. *Data mining: concepts and techniques*, Morgan kaufmann.

HICKS, B. J. 2007. Lean information management: Understanding and eliminating waste. *International Journal of Information Management,* 27**,** 233-249.

HIMMA, K. E. 2007. The concept of information overload: A preliminary step in understanding the nature of a harmful information-related condition. *Ethics and Information Technology,* 9**,** 259-272.

HJØRLAND, B. 2010. The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology,* 61**,** 217-237.

HOPKINS, D. J. & KING, G. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science,* 54**,** 229-247.

JACOBS, I. & WALSH, N. 2004. *Architecture of the World Wide Web, Volume One, W3C Recommendation* [Online].

KANTARDZIC, M. 2011. *Data mining : concepts, models, methods, and algorithms,* Hoboken, N.J., IEEE Press.

KELTON, K., FLEISCHMANN, K. R. & WALLACE, W. A. 2008. Trust in digital information. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,* 59**,** 363-374.

KENT, M. L., CARR, B. J., HUSTED, R. A. & POP, R. A. 2011. Learning web analytics: A tool for strategic communication. *Public Relations Review,* 37**,** 536-543.

KNIGHT, S.-A. & BURN, J. M. 2005. Developing a framework for assessing information quality on the World Wide Web. *Informing Science: International Journal of an Emerging Transdiscipline,* 8**,** 159-172.

KORDON, A. 2009. *Machine Learning: The Ghost in the Learning Machine,* Berlin, Heidelberg, Springer.

LANDIS, J. R. & KOCH, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159-174.

LANGFORD, L. 2010. Surf's up: harnessing information overload. *IEEE Engineering Management Review,* 38**,** 164-165.

MAI, J.-E. 2013. The quality and qualities of information. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,* 64**,** 675-688.

NTOULAS, A., NAJORK, M., MANASSE, M. & FETTERLY, D. 2006. Detecting spam web pages through content analysis. *Proceedings of the 15th international conference on World Wide Web.*

PALMER, J. W. 2002. Web site usability, design, and performance metrics. *Information systems research,* 13**,** 151-167.

PRIETO, V. M., ÁLVAREZ, M. & CACHEDA, F. 2013. SAAD, a content based Web Spam Analyzer and Detector. *Journal of Systems and Software,* 86**,** 2906-2918.

RYDER, M. 2005. *Semiotics: Language and Culture* [Online]. Detroit: Macmillan Reference USA. 4].

SAVOLAINEN, R. 2011. Judging the quality and credibility of information in Internet discussion forums. *Journal of the American Society for Information Science and Technology,* 62**,** 1243-1256.

SCHMIDT, N.-H., EREK, K., KOLBE, L. M. & ZARNEKOW, R. Towards a procedural model for sustainable information systems management. System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on, 2009. IEEE, 1-10.

SHARAPOV, R. V. & SHARAPOVA, E. V. 2011. Using of support vector machines for link spam detection. *INTERNATIONAL CONFERENCE ON GRAPHIC AND IMAGE PROCESSING (ICGIP 2011).* 1000 20TH ST, PO BOX 10, BELLINGHAM, WA 98227-0010 USA: SPIE-INT SOC OPTICAL ENGINEERING.

SHEN, G., GAO, B., LIU, T.-Y., FENG, G., SONG, S. & LI, H. 2006. Detecting link spam using temporal information. *ICDM 2006: SIXTH INTERNATIONAL CONFERENCE ON DATA MINING, PROCEEDINGS.* 10662 LOS VAQUEROS CIRCLE, PO BOX 3014, LOS ALAMITOS, CA 90720-1264 USA: IEEE COMPUTER SOC.

SHMUELI, G. 2010. To explain or to predict? *Statistical Science,* 25**,** 289-310.

SHMUELI, G. & KOPPIUS, O. R. 2011. PREDICTIVE ANALYTICS IN INFORMATION SYSTEMS RESEARCH. *MIS Quarterly,* 35**,** 553-572.

STAMPER, R. 1991. The Semiotic Framework for Information Systems Research. *Information Systems Research: Contemporary approaches & Emergent Traditions*.

STAMPER, R. 1996. An information systems profession to meet the challenge of the 2000s. *Systems Practice,* 9**,** 211-230.

TAYLOR, R. S. 1986. *Value-added processes in information systems.,* Norwood, Ablex.

TURBAN, E., SHARDA, R., DELEN, D. & KING, D. 2011. *Business Intelligence: A Managerial Approach,* New Jersey, Prentice Hall.

VAN DER SPOEL, S., VAN KEULEN, M. & AMRIT, C. 2013. Process prediction in noisy data sets: a case study in a Dutch hospital. *Data-Driven Process Discovery and Analysis.* Springer.

WANG, R. Y. & STRONG, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 5-33.

WANG, W., ZENG, G. & TANG, D. 2010. Using evidence based content trust model for spam detection. *Expert Systems with Applications,* 37**,** 5599-5606.

WIJNHOVEN, F. 2012. *Information services design : a design science approach for sustainable knowledge,* New York, Routledge.

WIJNHOVEN, F. & AMRIT, C. 2010. Evaluating the Applicability of a Use Value-Based File Retention Method. *Proceedings of SIGSVC Workshop***,** 10-118.

WIJNHOVEN, F., AMRIT, C. & DIETZ, P. 2014. Value-Based File Retention: File Attributes as File Value and Information Waste Indicators. *Journal of Data and Information Quality (JDIQ),* 4**,** 15.

WIJNHOVEN, F., DIETZ, P. & AMRIT, C. 2012. Information waste, the environment and human action: Concepts and research. *IFIP Advances in Information and Communication Technology,* 386 AICT**,** 134-142.

YANG, Z., CAI, S., ZHOU, Z. & ZHOU, N. 2005. Development and validation of an instrument to measure user perceived service quality of information presenting web portals. *Information & Management,* 42**,** 575-589.