Student.world: a solution for pragmatic waste?

Author: Lars Phillip Lehmann
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT. The internet has become crucial to the everyday life of most people and organisations. The amount of information accessible via the internet is rapidly increasing, so it becomes more and more difficult to find meaningful and useful Information. Information that is not meaningful to the receiver can be referred to as information waste. This paper examines different forms of information waste and how they can be identified. A special focus is laid on information that is subjectively useful to the receiver. Various tools for detecting different forms of information waste are drawn from literature and some new tools are added. The elaborated tools and criteria are compared to the ones used by the social media websites Facebook and student.world in order to identify a potential solution to information waste on the pragmatic layer.

Supervisors: Fons Wijnhoven and Chintan Amrit

Keywords
Information waste, web spam, waste detection, pragmatic layer, relevance of information, usefulness of information

5th IBA Bachelor Thesis Conference, July 2nd, 2015, Enschede, The Netherlands.

1. INTRODUCTION

The Internet has become indispensable and ubiquitous to the everyday life of most people. Every day, a lot of data is created and stored on servers that can be accessed via the World Wide Web. Due to the rapid advances in information technology in the recent years, there has been an exponential growth of data and information (Chui, Filbir, & Mhaskar, 2015).

In Web 2.0, users can create their own content, which leads to an increase of data generation and information flow. Thus, the Internet became a key component for sharing information and knowledge (Safko & Brake, 2009). Most organisations use the Internet and wireless communications in their businesses in order to communicate easily within the organisation, or with their customers (Chaffey, 2009). The Internet is mostly used for collaboration, but also education and entertainment are essential ways of using the Internet (Safko & Brake, 2009). Blogs and other interactive technologies on the World Wide Web were the first platforms for interactive information exchange. Later, they were almost completely replaced by social media, where everyone could contribute their knowledge, opinions, and other content (Golbeck, 2013).

The overwhelming amount of information created on the Internet and the Intranet of firms every day may include the creation of a lot of redundant and less valuable data. This data might hinder access to more valuable and relevant information, or at least make it more difficult to find. Since information and systems are critical for effective and efficient operations in organisations, waste has to be eliminated in order to ensure value flows (Hicks, 2007). The huge amount of unnecessary, redundant, and valueless data on the Internet is stored on many various servers, which consume a lot of energy in order to maintain these data. Their GHG emission can harm the environment, without adding value to Internet users (Wijnhoven, Dietz, & Amrit, 2012).

Web spam is created in order to manipulate search engine results and make it more difficult for users to find relevant data (Luckner, Gad, & Sobkowiak, 2014). The detection is still a problem to the Internet community, even though new spam filtering approaches seem to be promising (Filasiak, Grzenda, Luckner, & Zawistowski, 2014). The intention of spammers is often to get more visitors to their site, so they gain more money by advertising. Spam is also used for various criminal activities, like gathering personal information (Yu, 2015).

Also the quality of some material on the Internet might be questionable, since there are no consistent standards for publishing content on the World Wide Web or the Intranet. Qualitative problems include out-dated information, confusing layout and navigation, lack of consistency, reliability, security, accessibility, and more (Eppler & Muenzenmayer, 2014). Thus, there is a problem with the quality, as well as with the quantity of data and information. The latter is referred to as 'information overload' in literature (Himma, 2007). However, due to the varying forms of this content and material, it is difficult for detectors to identify.

In order to solve the problems of information waste and web spam, there has to be an effective and efficient filtering of data. In order to do so, these detectors need accurate filtering methods. There are already filters available, but the existing filters have difficulties to determine the value of a site or content to an individual user (Amrit, Wijnhoven, & Beckers, 2015).

This research paper focuses on the current topic of information waste and web spam. The goal of this paper is to review current methods and tools for filtering spam, and to make recommendations for dealing with information waste. A special focus is laid on the pragmatic layer, since this is missing in most existing approaches. The research question is: what are effective tools and criteria to identify information waste on the pragmatic layer on the Internet and Intranet?

1.1 Definitions

There are not many definitions of information waste in the existing literature, so it can be useful to look at definitions of information first. Information can be defined as a subset of data, which is meaningful to someone and can be considered as useful, significant, or urgent (Boddy, 2011). Everyone is depending on information, also organisations, where information can include cost and availability of resources or staff (Boddy, Boonstra, & Kennedy, 2008).

Contrary to information, information waste has no value to the receiver. The receiver perceives information as waste if data is unnecessary or unusable to him (Wijnhoven et al., 2012). This can be redundant data, or data that the receiver does not understand or trust.

Hicks (2007) describes waste as additional actions and any inactivity that arises if the receiver of information is not provided immediate access to the right amount of up-to-date information at the right time. In other literature, information waste is referred to as spam content, for example unwanted spam information and messages (Yevseyeva, Basto-Fernandes, Ruano-Ordás, & Méndez, 2013).

Literature distinguishes between information waste and web spam. Web spam is a specific form of information waste. The method web spam aims at manipulating search engine results by improving ranks of spam pages (Luckner et al., 2014). This includes all techniques that are used in order to get a high rank undeservedly. It is one of the main problems of the World Wide Web (Prieto, Álvarez, & Cacheda, 2013). The intention of Web Spam can be to earn money by gaining more traffic on the website, but also malicious behaviour is possible, like phishing. Phishing aims at acquiring user's private information, especially usernames, passwords, and social security numbers (Dhamija, Tygar, & Hearst, 2006). Spam pages can for example gather personal information about users to invade their privacy (Janssens, Nijsten, & Van Goolen, 2014).

Another consequence of spamming is that search engines become flooded with useless websites, which raises the cost of every search request (Gyöngyi & Garcia-Molina, 2005).

Thus, web spam makes up only a part of information waste, but an important one. A web spam filtering technique should be able to distinguish between spam and non-spam content.

Contrary to web spam, the intention of information waste is not always harmful, because it is often information that is relevant to some recipients, but irrelevant to others. For the latter, this can be classified as information waste on the semantic or pragmatic layer. These are two of the four layers of information waste, which are further described in the next section.

1.2 Layers of Information Waste

The layers of information waste are described in detail by Amrit et al. (2015). The identification of waste is more difficult in some layers than in others. These layers are the empiric layer, syntactic layer, semantic layer, and pragmatic layer.

Information waste on the empiric layer can be identified by detectable patterns that differentiate information waste from information. This means, if the message cannot be identified as such, it does not contain information. On this layer, the message is often considered as background noise (Boell & Cecez-Kecmanovic, 2010). Empirics are easy to identify for web spam filters since information waste on this layer is not dependent on the receiver.

Syntactics observe the representation of information and identifies content that is incomprehensible to the receiver (Beckers, 2014). For example when multiple different languages are used in one sentence on a site, the content might be information waste to most receivers. Also too many words on a page, in the page title or keyword stuffing are indicators for information waste on a syntactic layer.

Semantics identify whether a message is meaningful to the recipient. If a message cannot be integrated into the knowledge of the receiver, it will not be understood (Amrit et al., 2015) and can be classified as information waste on the semantic layer for this person. Also, too much, as well as too little specificity or detail of a topic is not useful to most recipients, since the content might not be understood or there might be an overload of information (Boell & Cecez-Kecmanovic, 2010). Information waste on this layer has no meaning to the recipient or aims to mislead the user. It is rather difficult for web spam filters to identify semantics of a website.

The pragmatic layer refers to the subjective usefulness for an individual to achieve his goal. Information is subjectively useful, if the receiver thinks the information can help him. If the information is not relevant in a specific situation, it can be classified as information waste for this individual. Also the novelty of information to the receiver is important, since providing him information that is already known might not be valuable to him, since it does not make him more informed and is redundant. This is not always the case because redundant information can be used for validation of existing knowledge (Boell & Cecez-Kecmanovic, 2010). Useful information on this layer is very situational and subjective for a specific individual. This makes it very difficult for spam filters to identify this kind of information for all users.

Contrary to the subjective usefulness that is an evidence for information waste on the pragmatic layer, objective usefulness refers to the actual usefulness of information. In this case, objectively useful information is necessary for better decision making, independently of what the receiver perceives as useful or helpful (Althuizen, Reichel, & Wierenga, 2012). Both, subjective and objective usefulness can be indicators for information waste on the pragmatic layer.

Most spam filters can identify information from the first two layers, but detection becomes more difficult in the semantic and especially the pragmatic layer. The reason for this is that relevant information on the empiric and syntactic layer is rather objective, while valuable information on the semantic and pragmatic layer is very subjective.

Metaphorically, the layers of information waste can be seen as the layers of an onion, as shown in Figure 1. Information waste on the syntactic layer includes the empiric layer, since background noise can usually not be understood by an individual. Since data that cannot be understood do not have much meaning to the receiver, information waste on the semantic layer includes the syntactic layer. Finally, information that has no meaning to the receiver is not considered as relevant and useful, thus the pragmatic layer includes the semantic layer, and usually the other layers as well.

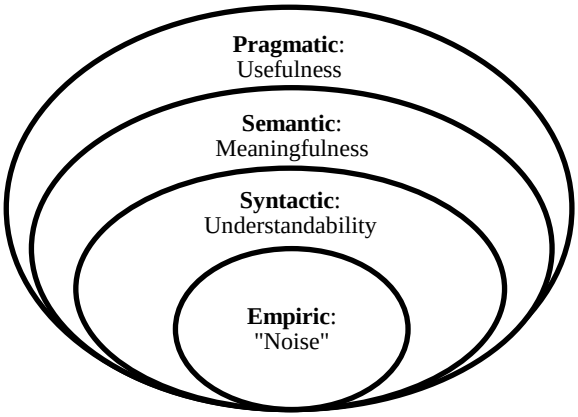

Figure 1. Four Layers of Information Waste.

1.3 Problem Statement

Since information waste is a current problem, there are already some approaches to solve it. However, these approaches do not solve the problem reliably yet. Especially since spammers always find new ways to circumvent a filter once it is established (Yevseyeva et al., 2013). For example, using image spam can circumvent a filter based on specific lexical terms. Thus, spam filters can become obsolete in short time, if they do not include appropriate learning algorithms.

Another problem is the different forms that information waste and web spam can have. As already mentioned, there are different layers of information waste. The information on these layers differs, as the ease of measurement on every layer does. Information waste on the syntactic layer can be in the form of unrelated links or keywords stuffed on a website in order to improve the rank of a site in search engine results (Prieto et al., 2013). Information waste on the semantic layer can be content that has no meaning to the recipient or it purposely misleads the recipient (Sharapov & Sharapova, 2011). The latter is more difficult to identify for web spam filters. It is even more difficult for web spam filters to recognise information waste on the pragmatic layer, since the usefulness of information differs for every individual in every situation. For example a cooking recipe for meatballs might not be valuable to a vegetarian.

Another difficulty that a filter faces is to report spam websites only, and not to mark valuable information as spam. The error when a non-spam website is filtered out as spam is referred to as false positive in literature (Yevseyeva et al., 2013).

A lot of things have to be considered when defining a filter for information waste. Thus, the goal of this paper is to identify the best possible filtering tools and criteria, which cover all layers of information waste.

2. CURRENT METHODS FOR IDENTIFYING INFORMATION WASTE

There are some spam filters in the current literature, which work in a different way, but have a similar goal; eliminating spam and information waste from the World Wide Web. The four filtering techniques and methods that will be reviewed are Spam Analyzer and Detector (SAAD), SpamAssassin, the content trust model, and optical character recognition (OCR).

2.1 SAAD

The SAAD is described by Prieto et al. (2013) as a multi-tiered system, which can decide whether a website can be classified as spam or not. It consists of a content analyser, a system configurator, a web spam decision tree, web spam repository, and a supervised result analyser. The content analyser is responsible to filter according to the heuristics, which accord to content analysis. The heuristics proposed by Ntoulas, Najork, Manasse, and Fetterly (2006) and Prieto et al. (2013) are among others, the number of words in the page, number of words in the page title, word length, ratio of the visible content, anchor text, compression ratio, number of common words, and specific phrases. The other heuristics that are used were not as effective in identifying web spam.

Number of words in the page includes keyword stuffing, which is a popular practice when creating spam pages. This means the website is extended by some popular words that have nothing to do with the rest of the site, thus becoming higher ranked by some search engines. Ntoulas et al. (2006) found that a page with more words has a higher probability of being spam. In total numbers, more than half of all web pages contain fewer than 300 words; only 12.7% of the websites contain more than 1000 words. However, there is a high rate of false positives, so there is a need for more heuristics.

Number of words in the page title is used as a heuristic, since some search engines focus on the content in the title of a website, so spammers use keyword stuffing also in the title of some spam pages. Ntoulas et al. (2006) show that some spam websites have important keywords in the title but useless content to a human viewer.

Keyword stuffing has changed over time. In a newer technique, words are written together in order to make longer ones (for example "freepictures"). So, word length has become a necessary and useful heuristic.

Ratio of the visible content refers to another way of keyword stuffing, which includes these keywords in not visible parts of a website, for example in comment in the HTML code.

A lot of search engines consider the anchor text of a link, which describes the content of the page. This is why spammers use an "anchor" to describe links that redirect to spam pages.

Spammers repeat some words or the whole content several times in order to improve the ranking of their pages. The level of redundancy can indicate web spam. This can be identified by the compression ratio which is calculated by the size of the normal website by the size of the compressed website.

Popular words of a specific language are added in order to be the response to many queries and show up in search results more often. So a web spam detector should count the number of common words.

Spam pages often contain common terms and specific phrases, like "drugs", "Viagra", "urgent", etc., which a web spam detector can easily identify.

The Spam Analyzer and Detector can identify many spam pages, but it also classifies some legitimate pages as spam (Prieto et al., 2013). The heuristics of the SAAD aim at reducing information waste on the syntactic and empiric layer. It already contains many important heuristics but it is not a complete model, since some tools are missing in order to reduce waste on the semantic and pragmatic layer.

2.2 Apache SpamAssassin

The SpamAssassin is described and tested by Méndez, Reboiro-Jato, Díaz, Díaz, and Fdez-Riverola (2012) and Yevseyeva et al. (2013). The SpamAssassin uses several techniques and plug-ins that allow setting up a personalised spam filter. Thus, it serves as the basis for other filters. Generally, SpamAssassin is used to filter spam e-mails with various spam filtering techniques, like searching for specific expressions over the content of the whole mail (Méndez et al., 2012). When new kinds of spam mails appear, the system administrator can simply add new rules or modify existing ones. Most of these tools can be used to identify information waste and web spam as well.

Examples of filter type techniques are Naïve Bayes, Language Guessing, DNS-based Blackhole List, SpamCop, and Content parsers. Naïve Bayes is a tool that calculates the probability of an e-mail being spam by analysing the text (Metsis, Androutsopoulos, & Paliouras, 2006). Language Guessing is a technique, which analyses the language an e-mail is written in and then identifies the comprehensibility for the receiver. The technique DNS-based Blackhole List saves Internet Protocol (IP) addresses of users who are responsible for spam and blocks their incoming e-mails. SpamCop is a service, where users can report spam e-mail senders. A similar tool could be used to report spam websites. Another technique is content parsers, which checks the correctness of the message body structure and the structure of HTML code in the e-mail (Yevseyeva et al., 2013), but it could also check for these on a website.

There are more possibilities to extend the usability of SpamAssassin with plug-ins and features to improve and customise its spam detection capabilities. SpamAssassin is highly customisable and configurable, so it can adapt to the users' personal needs. However, due to the subjective nature of e-mails, there cannot be a single filter that accurately classifies everybody's e-mails. This is also a reason why the SpamAssassin is prone to false positives. The SpamAssassin can identify spam on an empiric, syntactic, and even on the semantic layer, since it adapts to the users' needs. The SpamAssassin still has problems to identify information waste on the pragmatic layer because it cannot identify the subjective usefulness of a website.

2.3 Content Trust Model

The content trust model for spam detection is described by Wang, Zeng, and Tang (2010). The model helps to decide whether to trust a website based on its content. If a website cannot be trusted, it is most likely information waste or web spam. The content trust model distinguishes web pages in "good", "normal", and "spam". The websites are ranked via various evidences. Most evidences are independent from the language a web page is written in. The authors distinguish between text feature based evidence and information quality based evidence. Text feature based evidences include keyword stuffing, length of words, number of words in the page and page title, amount of anchor text, fraction of visible content, and amount of globally popular words. These heuristics were already described in the section about the SAAD. Additionally, Wang et al. (2010) propose to consider IP addresses, which are referred to by many symbolic host names, and the rate of evolution of web pages on a site.

Information quality based evidence has not been discussed before in the context of spam detection. Wang et al. (2010) identify six evidences which aim to evaluate the quality of a website automatically. These are currency, availability, information-to-noise ratio, authority, popularity, and

cohesiveness. Currency indicates whether a website has been updated recently. Out-dated websites do rarely contain valuable information. Availability refers to the amount of broken links on the website, which influences the meaningfulness of that site. The information-to-noise ratio is the ratio between meaningful information and size of a website. This evidence is important, since valuable information might be difficult to find on a website with a lot of content. Authority is the reputation of the organisation behind that website, based on the Yahoo Internet Life reviews. This can be useful to prioritise more trustworthy content. Popularity is measured by the number of other websites that have cited a certain website in order to assess its usefulness. Finally, cohesiveness is the degree to which a website focuses on one topic. Spam websites tend to focus either too much on one topic or they have many topics that are not related at all.

These evidences provide efficient and rather accurate techniques that can identify most of the web spam and have a low amount of errors in the tests. Especially the information quality based evidences are novel and helpful in identifying spam. The Content Trust Model is able to identify web spam and information waste on an empiric, syntactic, and semantic level. However, since this model is not customised to specific users, it has problems to identify information waste on a pragmatic layer. This model focuses more on the reliable detection of text and information quality based web spam, instead of the users' needs.

2.4 Optical Character Recognition (OCR)

The previous tools and methods primarily identified text based spam and did not pay too much attention to image spam. Image spam is a prevalent form of spam that is important to consider. Image spam means that the spam message is embedded in images that are sent as mail attachments, in order to circumvent text based spam detectors (Biggio, Fumera, Pillai, & Roli, 2011). OCR-based techniques can analyse text embedded in attached images. The techniques keyword detection and text categorisation are used. Even though this method is used to identify e-mail spam, it can also be used to identify information waste and web spam on websites.

Keyword detection refers to identifying spam in e-mails by checking for typical keywords that often appear in spam e-mails. Spammers can easily circumvent this by misspelling those words. However OCR can be used to extend the previously mentioned SpamAssassin, so that spam images in e-mails can be identified and additionally users can extend the keyword list in order to have more possibilities for customisation. There is also an existing plug-in available that filters out messages that contain a high amount of misspellings. Text categorisation aims to improve image spam detection rate based on machine learning and pattern recognition (Fumera, Pillai, & Roli, 2006).

OCR-based techniques generally have a low false positive rate, but also a rather low true positive rate, since all images without text were classified as no spam. These techniques usually detect information waste on the empirical and possibly on the syntactic layer only, but information waste on the semantic and pragmatic layers is neglected. The reason is that OCR tools only focus on whether an image in an e-mail attachment contains text that is usual for spam, but it does not consider how meaningful and valuable that information might be to the receiver. OCR does not deliver a complete set of tools for identifying information waste but its tools can be used in addition to other spam filter tools that cannot detect text in images.

3. DETECTING INFORMATION WASTE ON THE PRAGMATIC LAYER

Since none of the mentioned filters can detect information waste on the pragmatic layer reliably, criteria of how such a tool could work are described in this section.

Pragmatic information has to be novel, if the recipient does not need validation, relevant for the individual's goal and situation, valuable to the recipient, trustworthy, accessible, and timely (Boell & Cecez-Kecmanovic, 2010).

The novelty of information can be ensured if the individual is not shown the same article or document more than once, except if he needs confirmation of his knowledge. In the best case the recipient can receive similar information from another website in order to validate his knowledge. However, in most cases it can be useful for the recipient to revisit a familiar website for the same information. Since this evidence is debatable, its priority is low and can be neglected for now.

Goal relevance is important since information has to be useful to the recipient in the current situation in order to achieve a specific goal. The goal relevance corresponds to the subjective usefulness. The tool has to learn about the needs of the user by his queries and other information it can get. Also, the tool could consider the time, date, and place for specific queries. For example if the user searches for a restaurant in the evening, the tool should put emphasis on restaurants that are open at the moment and nearby.

Value to the recipient is similar to goal relevance, but distinguishes itself by its instructional and economic value. The value is independent of the perception of the recipient, so it is objectively useful. Information with instructional value helps users to make decisions or solve problems, while economic value helps users to make profit or avoid costs. A tool should be able to categorise every website and so assess these values for their users by using metadata ("data about data").

At some point in time, data might be valuable information, but become irrelevant at another time. For example, information about mergers and acquisitions are relevant when they are released, but their usefulness declines fast (Boell & Cecez-Kecmanovic, 2010). This time dependence can be transferred into a tool, which prefers recent events over older ones.

As stated by some authors (Du & Arif, 2011), (Savolainen, 2011), (Wang et al., 2010), trustworthiness of information is also an important attribute of information quality. Trustworthiness is also referred to as credibility of Information. Users only act on information they trust, thus, information that users do not trust, is useless to them. A tool, which shows only trustworthy sites, should be able to check the author and creator of a website or article and identify his credibility. Black lists should be created and users have to be able to contribute to the list by reporting creators of suspicious content.

From my own experience, information is only useful, if it is accessible. The user does not have benefits if a search engine shows him a site that he does not have access to. When identifying useful information, a tool has to check whether the user has permission to enter a website, read a document, watch a video, or else. According to the definitions for each layer, the tools that check the trustworthiness and the accessibility of a website rather serve to identify information waste on the semantic layer.

In summary, a method to identify pragmatic information waste has to be customisable for every user, be adaptive, and use metadata to assess value and the timeliness of data. The credibility of a content creator and the accessibility of content will be added to the tools for identifying information waste on the semantic layer. These evidences can be used in addition to the ones for the empiric, syntactic, and semantic layers of information waste. In table 1, a short overview of the most important methods and tools is shown for each layer of information waste.

3.1 Student.world

Student.world is a social platform designed by and for students. On the website, information is shared and ranked by students. The goal is that the user only receives information that is relevant for him, instead of irrelevant stories and spam. Student.world is specifically for students and the website additionally learns about the interests of users, so the information is very specific for every user. Spamming is not allowed and new content is reviewed by other users first, so it is difficult to distribute spam on the website. Also, student.world incorporates a level system, which means that users are only allowed to add new items, move items, or vote, once they reached a certain level by spending time on the website or other actions (Korevaar, 2015).

Student.world is not a typical spam filter or detector; however the website serves as an example of how social platforms for private persons and organisations could work in order to avoid spam and information waste. The website might look a bit confusing and overwhelming at first, due to the hierarchical scheme of categorisation (Figure 1), but it is actually easy to navigate and fast to learn. Also the search function on the website is useful to find specific information fast. The question is: can student.world help in avoiding information waste on the pragmatic layer for its users? In order to answer this question, the tools for identifying information waste are compared with the features of student.world. The website tries to give the user only information that he is interested in. If the website evaluated that a user is interested in information waste and web spam, the website shows these to the user. This means that student.world does not necessarily filter out information waste on the empiric and syntactic layer.

Student.world works on the semantic layer, because it shows more current and available information. Links to content that is not available anymore, also known as dead links, will be deleted by the community. Authority is ensured, because the community is only meant to share content that can be trusted and whose authors have a good reputation. The popularity of content on student.world does not work as explained by Wang et al. (2010). On student.world the users can show that they like a topic or content by giving a kudo and optionally describing what they think is good about it. If they do not like something, they can take a kudo away. The popularity of an item can be measured by the amount of kudos and views. More popular content is shown first to a user. The content of the website is not too narrow on a single topic, since there are many topics. Neither is the content too far spread, since all topics are relevant to certain students. A user is able to receive a sufficient amount of detail about many topics, which means that cohesiveness is also given on student.world. Information-to-noise ratio can be reduced with every step further into the hierarchy of a topic. Generally, there is not much noise, since all information there can be seen as a message and is understandable by someone. Student.world can assess the criteria by Wang et al. (2010), which are currency, availability, information-to-noise ratio, authority, popularity, and cohesiveness. This means that student.world is able to reduce information waste on the semantic layer.

Table 1. Tools for identifying information waste on each layer.

Layers	Criteria	Tools	Methods
Empiric	- Language skills of the user - Block authors known for spam content - Recognition of specific patterns	- Language guessing - DNS-based Blackhole List - SpamCop - Keyword detection - Text categorisation	SpamAssassin (Méndez et al., 2012); (Yevseyeva et al., 2013) OCR Tools (Biggio et al., 2011); (Fumera et al., 2006)
Syntactic	- Keyword stuffing - High amount of invisible text - Specific words and phrases - Repetition of words or content - Anchor words	- Words per page/ page title - Word length - Ratio of visible content - Ratio of anchor words - Compression ratio - Ratio of globally popular words	SAAD (Prieto et al., 2013); (Ntoulas et al., 2006) Content Trust Model (Wang et al., 2010)
Semantic	Information quality based features: - Currency of information - Authority of the author - Popularity of content - Availability and accessibility of content - Cohesiveness	- Time stamp of the last modification - Ratio of broken links on a page - Yahoo Internet Life reviews - Number of links to the site - Information-to-noise ratio - Degree of relation of major topics	Content Trust Model (Wang et al., 2010)
Pragmatic	- Adaptability to the needs of the user - Customisable to achieve a relevant goal - Assessment of instructional and economic value - Time dependency	- User profiles - Consideration of time and place - Metadata - Preference of more recent data	

The tools that were elaborated for identifying information waste on the pragmatic layer are customisability, adaptability to the user needs, assessment of instructional and economic value, and prioritisation of recent data. The information that the user sees on student.world can be influenced by the topics or users they follow. This customisability helps a user to achieve his goal faster, since more relevant information is shown in his feed. It is more difficult to identify whether the website can show information with economic and instructional value. However, no specific tool for this purpose could be identified on student.world, so this has to be improved if they want to deliver more value to the users. The last tool is the prioritisation of more recent data and this is what student.world is already doing in their news feed. Users are also able to choose between newly added content, content that was popular last day or last week, and the most popular content of all time on the website.
Student.world fits some criteria that were made for identifying information waste on the pragmatic layer, so it can be a possible solution to reduce this waste. However, the assessments of instructional and economic value have to be added.

At the moment, the platform is for students only, but if the project works well, some platforms for other groups might be established as well. Another possible application of the student.world structure will be described in the next section.

3.1.1 Possible Application

Since student.world seems to be able to solve the problem of information waste on the pragmatic layer to a certain degree, the principle might be useful in other areas, for example for organisational infrastructures, especially intranets. Data exchange within companies is sometimes difficult and unclear, as I experienced during my internship at Greif. Greif Germany GmbH is a company that produces and sells all kinds of industrial packaging. Quick data exchange and communication is very important in this business, since orders have to be dispatched fast and neatly. Currently, the datasheets for their products are on their local server and can be accessed via the Windows browser. However, this method takes a long time since there are a lot of different products and their specifications, which are not always labelled correctly or in the designated folder. To find datasheets faster, a hierarchy like the one of student.world could be established to manage them. This method can also be used for invoices, contracts, dispatch notes, and other files and documents that employees need quick access to. If everything is neatly organised within one program, there is less noise and information overload, and employees are able to find information that is valuable and relevant to them. Since the organisation often needs to ask for data and information from other branches, they could establish a hierarchical structure for their files in more branches and make them accessible for all employees with permission. Thus, using the hierarchical approach for structuring files could reduce information waste in organisations.

3.2 Facebook

Facebook is another social media platform, where users can create and share content. Contrary to student.world, users do not vote on content directly. However, a user can like content to see more similar content in the future, or report that he does not want to see this post and less similar posts in the future. The goal of Facebook is to show the user timely, relevant, trustworthy, high quality content in their news feed (Kacholia, 2013). The news feed is individual for every user and influenced by their interests, the interests of their friends, and the aforementioned factors. The web site also uses personal information to show the user targeted advertisement.

The website is adapting to the needs of every user and helps him to find relevant information by letting him create a user profile and add all kinds of data about himself; like age, gender, and interests. Based on this data, the website tries to identify what is relevant to the user.

The language of the content that the user sees is based on the

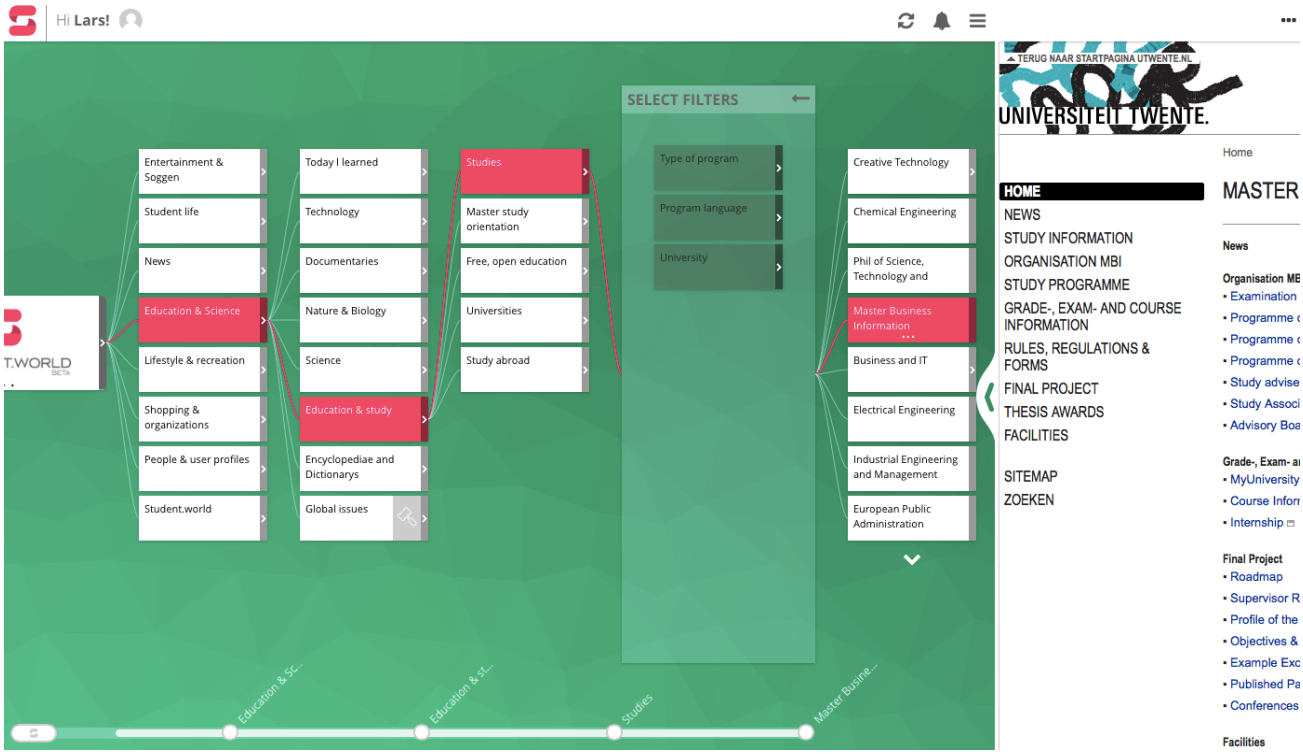

Figure 2. Example of a hierarchical search on student.world

language he chooses for the user interface, his interests and the languages his friends use, so there is no evidence of Facebook actively filtering content based on the language.

Users can report spam content and even what they perceive as information waste to the system, so that they see less similar content in the future. So, like at student.world, users can theoretically receive information waste on the empiric and syntactic layer, if they want to.

Wang et al. (2010) propose six features to assess the quality of information; currency, availability, authority, popularity, and cohesiveness. Facebook does not generally filter any content that does not match these criteria. However, the website uses an algorithm in order to show high quality posts higher in the news feed (Kacholia, 2013). One criterion of the algorithm is to prefer timely posts, which ensures that more current data is shown. Facebook does also consider the relevance of a post to users. The more interactions a user has with a friend or page, the more likely he is to see their posts. So, the authority of the author is also considered. Furthermore, more popular posts among Facebook users are more likely to show up in news feeds of other users, so it is ensured that the factor popularity is considered. Content that is removed from Facebook cannot be seen by users, so availability is given. The cohesiveness of information is subjectively dependent to every user based on his interests and behaviour on the website, so it is not always present. Overall, the algorithms at Facebook cover the semantic layer of information waste sufficiently.

On the pragmatic layer, adaptability to the needs of the user is important, which is usually given on Facebook with the personalised user profiles. With the information present about every user, the website can identify what is subjectively relevant to the user.

In order to achieve objective relevance, the website should be able to identify the goal of every user in a specific situation. However, this is not the actual purpose of the website. According to their initial public offering statement, the purpose of the website is to let people relate to each other and share content (Ebersman, 2012). Thus, the website does not actively help a user to achieve any goal, but does help the user if his goal is to connect with other people. This means the value of the website is depending on the situation. There is also no evidence that Facebook can assess the instructional or economic value of content, probably due to the same reason.

In summary, Facebook is able to identify information waste on the semantic layer reliably, and does a good job identifying and ranking valuable information on the pragmatic layer for the user.

3.3 Discussion

The goal of this paper is to identify the best possible filtering techniques for information waste. Several methods and evidences were identified, which can filter information waste reliably on the empiric and syntactic layer. Some methods are able to filter information waste on the semantic layer but it is especially difficult to filter information waste on the pragmatic layer. Tools to identify information waste on the pragmatic layer were not sufficiently covered in literature, so some tools for this purpose were proposed. These tools were compared to the ones used by the websites student.world and Facebook. These websites are not web spam filters or search engines themselves, but they show how filters should work in order to detect information waste on the pragmatic layer. Specifically, student.world was chosen because it is very specific and a rather new approach to social media. It has a unique interface and is strongly dependent on input by users. Content is not created by engines and neither are the user profiles, so users have full control over their profile and the content they see. Even though there is not much content on student.world yet, the concept is good in order to avoid information waste on the Internet. Facebook was chosen since it is very popular and aims at showing relevant information to the user first. The comparison showed that student.world and Facebook are already doing a good job filtering and sorting information waste on the pragmatic layer because they deliver personalised results based on the interest of the user and can thus estimate the subjective usefulness of information. However the websites lack to identify economic and instructional value of content. These websites may use more tools than the ones mentioned, but those are not relevant for this research.

For further theoretical research, these results deliver important tools for filtering information waste on the pragmatic layer. Further theoretical research, more examples besides student.world and Facebook could be identified and analysed for their pragmatic value.

Since this paper only focuses on the theoretical identification of practical methods, the quality of evidence is rather limited. Thus, in further research the elaborated tools and criteria could be tested to confirm or reject whether they can reduce information waste on the pragmatic layer in practice.

4. CONCLUSION

On the Internet, users often face problems when they search for useful and relevant information, due to the high amount of information waste and web spam. There are already several approaches to the elimination of information waste and web spam, and some are more effective for different kinds of information waste than others. Most existing literature only focused on identifying web spam and information waste on the empiric, syntactic, and semantic layer, while the pragmatic layer was often not or just barely considered. This research paper focused on this often neglected layer of information waste. The research question of this paper is: what are effective tools to identify information waste on the pragmatic layer on the Internet and Intranet? Some criteria were proposed that are able identify information waste on the pragmatic layer. These tools are: customisability to a relevant goal considering circumstances, adaptability to the needs of the user, categorisation of websites and assessment of instructional and economic value, and the prioritisation of recent data to ensure users get more useful information. Based on these some tools were identified that can check if these criteria are met. User profiles can help to get to know a user and his interests and thus, adapt to his needs. When time, place, and other circumstances are considered, a filter should be able to help a user achieve a relevant goal. Metadata of a website can be helpful for a filter in order to identify the instructional and economic value of content. Finally, when more recent data is prioritised, the possibility is higher that a user receives novel and timely relevant information. These criteria are similar to the ones that are often found on social media websites. So, another question that was asked in this research paper is: can student.world help in avoiding information waste on the pragmatic layer for its users? Student.world is able to use most of these tools; however there is some room for improvement. The criteria were also compared to the algorithm that Facebook uses to sort its content. The results were similar. According to the results of this research, social media in general could deliver a solution to pragmatic waste.

Overall, there is too little attention paid to information waste on the pragmatic layer in literature and practice, even though Internet users and employees in companies waste a lot of time searching for useful information. There has to be more research about this topic in order to shed more light on the topic of information waste and web spam.

5. ACKNOWLEDGEMENTS

I would like express my gratitude to my supervisors from the track of Business Information Management, Fons Wijnhoven and Chintan Amrit, for their assistance during this project, as well as their insightful research on this topic. I would also like to thank the student who wrote his bachelor thesis about the topic of information waste and web spam last year, David Beckers, for his useful work. Finally, I would like to credit Wim Korevaar and the student.world team for their work on the website that made an eligible case for this research.

6. REFERENCES

1. Althuizen, N., Reichel, A., & Wierenga, B. (2012). Help that is not recognized: Harmful neglect of decision support systems. *Decision Support Systems, 54*(1), 719-728. doi: 10.1016/j.dss.2012.08.016
2. Amrit, C., Wijnhoven, F., & Beckers, D. (2015). *Information Waste on the World Wide Web and combating the clutter*. Paper presented at the Twenty-Third European Conference on Information Systems (ECIS), Münster, Germany.
3. Beckers, D. (2014). *Information waste on the World Wide Web: combating the clutter*. Paper presented at the 3rd IBA Bachelor Thesis Conference, Enschede.
4. Biggio, B., Fumera, G., Pillai, I., & Roli, F. (2011). A survey and experimental evaluation of image spam filtering techniques. *Pattern Recognition Letters, 32*(10), 1436-1446. doi: 10.1016/j.patrec.2011.03.022
5. Boddy, D. (2011). *Management: An Introduction* (Vol. 5). Harlow: Financial Times Prentice Hall.
6. Boddy, D., Boonstra, A., & Kennedy, G. (2008). *Managing Information Systems: Strategy and Organisation*. Harlow: Prentice Hall/Financial Times.
7. Boell, S., & Cecez-Kecmanovic, D. (2010). *Attributes of Information*. Paper presented at the AMCIS 2010 Proceedings, Lima, Peru. http://aisel.aisnet.org/amcis2010/129
8. Chaffey, D. (2009). *e-Business and e-Commerce Management: Strategy, Implementation and Practice* (4 ed.). Harlow, England: FT Prentice Hall.
9. Chui, C. K., Filbir, F., & Mhaskar, H. N. (2015). Representation of functions on big data: Graphs and trees. *Applied and Computational Harmonic Analysis, 38*(3), 489-509. doi: 10.1016/j.acha.2014.06.006
10. Dhamija, R., Tygar, J. D., & Hearst, M. (2006). *Why Phishing Works*. Paper presented at the CHI-2006: Conference on Human Factors in Computing Systems, Quebec, Canada.
11. Du, J. T., & Arif, A. S. M. (2011). *Judgment of Information Quality during Information Seeking and Use in the Workplace: A Case Study of Marketing Professional*. Paper presented at the Proceedings of the 16th International Conference on Information Quality, Adelaide, Australia.
12. Ebersman, D. A. (2012). *Registration Statement, Facebook, Inc.* Washington, D.C.: Retrieved from http://www.sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm.
13. Eppler, M. J., & Muenzenmayer, P. (2014). *Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology*. Paper presented at the Seventh International Conference on Information Quality, Cambridge, MA.
14. Filasiak, R., Grzenda, M., Luckner, M., & Zawistowski, P. (2014). On the testing of network cyber threat detection methods on spam example. *annals of telecommunications - annales des télécommunications, 69*(7-8), 363-377. doi: 10.1007/s12243-013-0412-5
15. Fumera, G., Pillai, I., & Roli, F. (2006). Spam Filtering Based On The Analysis Of Text Information Embedded Into Images. *Journal Of Machine Learning Research, 7*, 2699-2720.
16. Golbeck, J. (2013). *Analyzing the Social Web*. Amsterdam: Elsevier MK.
17. Gyöngyi, Z., & Garcia-Molina, H. (2005). *Web Spam Taxonomy*. Paper presented at the Proc. 1st Int. Workshop on Adversarial Information Retrieval on the Web.
18. Hicks, B. J. (2007). Lean information management: Understanding and eliminating waste. *International Journal of Information Management, 27*(4), 233-249. doi: http://dx.doi.org/10.1016/j.ijinfomgt.2006.12.001
19. Himma, K. E. (2007). A Preliminary Step in Unterstanding the Nature of a Harmful Information-Related Condition: An Analysis of the Concept of Information Overload. *Ethics and Information Technology, 9*(4), 259-272.
20. Janssens, K., Nijsten, N., & Van Goolen, R. (2014). Spam and Marketing Communications. *Procedia Economics and Finance, 12*, 265-272. doi: 10.1016/s2212-5671(14)00344-x
21. Kacholia, V. (2013). News Feed FYI: Showing More High Quality Content. Retrieved 18.06.2015, from https://www.facebook.com/business/news/News-Feed-FYI-Showing-More-High-Quality-Content
22. Korevaar, W. (2015). student.world. Retrieved 19.05.2015, 2015, from http://www.student.world/?ref=418886
23. Luckner, M., Gad, M., & Sobkowiak, P. (2014). Stable web spam detection using features based on lexical items. *Computers & Security, 46*, 79-93. doi: 10.1016/j.cose.2014.07.006
24. Méndez, J. R., Reboiro-Jato, M., Díaz, F., Díaz, E., & Fdez-Riverola, F. (2012). Grindstone4Spam: An optimization toolkit for boosting e-mail classification. *Journal of Systems and Software, 85*(12), 2909-2920. doi: 10.1016/j.jss.2012.06.027
25. Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). *Spam Filtering with Naive Bayes – Which Naive Bayes?* Paper presented at the CEAS 2006 - The Third Conference on Email and Anti-Spam. http://www.aueb.gr/users/ion/docs/ceas2006_paper.pdf
26. Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006). *Detecting Spam Web Pages through Content Analysis*. Paper presented at the Proceedings of the World Wide Web conference.
27. Prieto, V. M., Álvarez, M., & Cacheda, F. (2013). SAAD, a content based Web Spam Analyser and Detector. *The Journal of Systems and Software, 86*, 2906-2918. doi: 10.1016/j.jss.2013.07.007
28. Safko, L., & Brake, D. K. (2009). Introduction *The Social Media Bible: Tactics, tools, and strategies for business success*. Hoboken, NJ: John Wiley & Sons.
29. 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*(7), 1243-1256. doi: 10.1002/asi.21546
30. Sharapov, R. V., & Sharapova, E. V. (2011). *Using of support vector machines for link spam detection*. Paper presented at the Proc. SPIE 8285 International Conference on Graphic and Image Processing (ICGIP 2011), Cairo, Egypt.
31. Wang, W., Zeng, G., & Tang, D. (2010). Using evidence based content trust model for spam detection. *Expert Systems with Applications, 37*(8), 5599-5606. doi: 10.1016/j.eswa.2010.02.053
32. Wijnhoven, F., Dietz, P., & Amrit, C. (2012). Information Waste, the Environement and Human Action: Concepts and Research. *IFIP Advances in Information and Communication Technology, 386*, 134-142.
33. Yevseyeva, I., Basto-Fernandes, V., Ruano-Ordás, D., & Méndez, J. R. (2013). Optimising anti-spam filters with

evolutionary algorithms. *Expert Systems with Applications*, *40*(10), 4010-4021. doi: 10.1016/j.eswa.2013.01.008

34. Yu, S. (2015). Covert communication by means of email spam: A challenge for digital investigation. *Digital Investigation, 13*, 72-79. doi: 10.1016/j.diin.2015.04.003