UPTEC STS08 031
Examensarbete 30 hp
December 2008

Finding representations of people from web appearances

Carl-Henrik Arvidsson

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Finding representations of people from web appearances

Finding representations of people from web appearances

Carl-Henrik Arvidsson

People are today, to various extents, represented by information put on the World Wide Web. For a single person, it is difficult to get an overview of how accurate this representation is. What is the image that emerges if one uses a search engine to extract information from the Internet about oneself? How can we illustrate an individual's web presence? A tool that illustrates, and allows us to explore online representations of identity, helps us to more fully understand the structure of online information related to individuals. The work presented here explores representations of people on the Internet. This is done through a manual study of web sites to find different dimensions of people who are represented on the pages in question. A dimension that was found was "sphere". Depending on the context different facets of an individual's identity are revealed.
One way to find representations of these dimensions from information on web pages was to analyze the text on the web pages using classifiers. The dimension "sphere" can be observed and classified by the techniques used in the study. In order to possibly refine the results the text was examined with different linguistic measurements to see how they could form the basis for classification.

Handledare: Markus Bylund
Ämnesgranskare: Arnold Pears
Examinator: Elísabet Andresdóttir
ISSN: 1650-8319, UPTEC STS08 031
Sponsor: SICS, Swedish Institute of Computer Science

Populärvetenskaplig sammanfattning

Mängden information på Internet växer explosionsartat och det ligger nära till hands att tänka att så även gäller information om individer på Internet. För en enskild person är det svårt att få en överblick över hur denne är representerad på Internet. En sökning på ens eget namn i en sökmotor, vilket bild ger det av en som individ? Ett verktyg för att kunna åskådliggöra hur en individ uppfattas av utomstående personer via den information som finns på nätet är en idé som kan vara av intresse för många. Detta arbete är en undersökning för att hitta representationer av människor på Internet. Detta görs genom en manuell grundlig undersökning av webbsidor för att hitta olika dimensioner av människor som är representerade på sidorna i fråga. En dimension som hittades var ”sfärer”, beroende på vilket sammanhang en individ vistas i framställer man olika aspekter av sin person. Vissa sidor kommer fram i sammanhang kopplade till jobbet, andra tillsammans med människor man känner privat en tredje sida är en ”publik” sida. Den kan exempelvis visas mot individer som är okända i förväg och där man har någon form av information man vill förmedla.

En väg mot att hitta representationer av de funna dimensionerna, där ”sfärer” är en, på webbsidor var att analysera textmassan på webbsidorna genom klassificeringsalgoritmer. De klassificeringar och mätningar som görs ger godtagbara resultat. För att möjligen kunna förfina resultatet undersöktes texten med olika lingvistiska mätningar för att se hur resultatet från dessa kunde ligga till grund för klassificeringar. Tillsammans med det textmaterial som användes i denna undersökning kan man tänka sig att använda bilder och layout som grund för ytterligare analys.

Acknowledgements

This report is the result of a master thesis project conducted at Swedish Institute of Computer Science, SICS, during the winter and spring of 2007/2008. Several people have been involved in the work leading to this report and deserve my gratitude. At SICS among many, a special thank you to my co-worker Pedro Sanches and my supervisor Markus Bylund. At home, my always supportive and encouraging Maija. Thank you all.

Uppsala, June 2008
Carl-Henrik Arvidsson

Table of contents

1. Introduction

In this day and age a lot of personal information is stored digitally. High distribution and large amounts of information gives a poor overview. Information about individuals that is gathered and stored by information and communication technology (ICT) systems is an important issue in the privacy discourse. May it be information from credit card usage, logs from search engines or which base stations a mobile phone passes by, revealing where a particular person is geographically situated. Most of this information is stored at the provider for each service and thus the information is not available for a comprehensive analysis. This also makes it harder for the individual to be aware of what information about her that is available. Information that is presented on the World Wide Web about a person, on the other hand, is publicly available for everybody to look at. Individuals can't be sure of what information about them is available. A feature of web information is that it is to be considered to stay there more or less forever once it gets online. When the information is present on the web it is impossible to know where it is spread and where that information can appear in the future. The information is no longer in the control of the person who published it on the web in the first place. An example of this is if a person uploads a photo to the photo sharing website Flickr® anyone viewing it can save a copy of that photo and republish it wherever she wants. It doesn't matter if the original publisher removes the photo, that person can't know where the photo might appear again. Another example is personal communication via email or instant messaging that often are offered to no obvious cost to the end user, information that the user expresses in one day and situation by these means could come up to light long time after in a different setting even if it is forgotten by the user in the first place. An individual does not have the power to know what information about her that is available, to be "safe" all information shared via information and communication technologies must be treated as publicly available. It is not practically doable for most people to keep track of what she shares with help of ICT systems; therefore there exists an *information imbalance* between the provider of the service and the end user. (Bylund et. al. 2008)

Information about an individual on the web can be of different kinds and be scattered to a large extent. The same piece of information can occur at several different places, the context of some information may tell us just as much or sometimes more than the information itself. Each part of this information and its origin reveals some pieces about the individual and facets of the person's identity are shaped. Though, there are a lot of uncertainties about the quality of the facets shaped. The problem to get relevant knowledge from personal information lays in the large amount of information scattered on the web and the difficulties for the individual to get a comprehensive view that is needed to be able to judge the quality of the facets presented of her. (Bylund et. al. 2008)

Do the facets really reflect the actual circumstances of a person's real life? How does other people perceive her when they find information related to her online? What facets of her identity are revealed for example when a recruiter does a web search for her name? How well do these facets represent her real identity? Questions like these may come to mind when thinking about personal information in combination with information and communication technologies. In this report answers to these questions won't be given, but what will be presented is an investigation of finding information

about individuals from web material. The investigation is used to provide support for a thesis:

By combining information extraction, natural language processing and data mining techniques it is possible to create representations of different facets of individuals' identities, as they appear on the World Wide Web.

Some means that enables someone to get an overview of the information about an individual on the web is an idea that would interest a lot of people. Who better to judge the accuracy of the facets the web provides than the person in question? In this report an approach for finding representations of facets of identities of people on the World Wide Web are presented. The representations can be used to address the information imbalance between described by enhancing individuals' awareness of their appearance on the web. As discussed, the web is one source of information among many. The web is chosen, as it is very accessible and open for analysis.

1.1 Outline

This report consists of two sections. The first section, chapter 1, 2 and 3 is an introduction to the subject, description of the theoretical background of the concepts and tools used in this report. The second part describes and discusses the implementation of a system that could be used for providing a comprehensive view of an individual's appearance on the World Wide Web. Chapter 4 describes the requirements and functions of that system. Chapter 4 also contains a case study conducted to test methods that possibly would be of use for the tasks to obtain data from the web and analyze it with regard to how individuals are presented. Chapter 5 discusses the results presented in chapter 4. At the end an appendix shows exerts from the implementation work for interested readers.

2. Methodology

To support the thesis a combination of tasks will be done. This work is partly based on a literature study, partly based on experimental work conducted to find and evaluate representations discussed. The implementation experiments will be done in an iterative way to be able to make use of the knowledge that comes to hand in the early stages of that process.

3. Theoretical foundations

3.1 Identity

Jenkins (2003) summaries two criteria for the notion of identity; similarity and difference. What people typically do when identifying something is to classify things or persons or associate one with, or attach oneself to, something or someone else. These criterions are the baseline for the concept of identity. Mead (1934, in Jenkins 2003) suggests that it is impossible for an individual to at all see her self without also seeing her self as other people see her. This is of course not easy to fulfill as it is impossible to step outside the body and observe one self. One's identity is traditionally tightly connected to the physical appearance; "Identification in isolation from embodiment is unimaginable" (Mead 1934, in Jenkins 2003) illustrates this traditional perception of identity. This view is not possible in the world of today. The modern world is full of non-body identification, especially with regard to information and communication technology systems. This difference in view between traditional theoretical concepts and the actual circumstances of today illustrates potential problems that the use of the concept identity can reveal.

3.1.1 Facets of identity

People engaging in social interaction typically don't disclose their full identity (Boyd, 2002). Individuals have developed mechanisms for maintaining a faceted social identity by controlling who has access to what parts of the personal information a person has available. The facets can be seen as parts of an individual's full identity. She usually reveals parts about her that makes sense and is useful in a particular situation. An example of this behaviour is that normally an individual don't immediately tell persons that they meet in a situation related to one's job about engagements with the children's football club or where the family vent for vacation last summer. Of course may this kind of information be shared at a later stage but then after conscious decisions.

3.1.2 Identity in the light of information and communication technologies

Boyd (2002) describes how these facets are represented in the digital world when communicating using information and communication technology (ICT) systems. When people are communicating via ICT-systems (e-mail, instant messaging, blogs etc.) they don't have the tools to present and manage their identity as they do when communicating face to face. This is an issue that must be accordingly managed.

3.1.3 Privacy as means for identity management

Identity management is a complex issue. The notion of privacy is connected to identity. The sociologist Irwin Altman defines privacy as: "selective control of access to the self or to one's group" (Altman 1975). A key word here is "selective", the control changes over time and between different social situations. Selective access is just what revealing facets of identity are about. Being able to keep the facets of one's identity separated is the basis of privacy. The level and perception of privacy is therefore a useful tool for identity management. As an individual regulate her privacy she also manages her identities. There are different ways to regulate privacy. Altman (1977) investigates how privacy is regulated in different societies around the world. He argues that the need and

the ability to regulate privacy is universal but the mechanisms for doing it differs and also the desired level of privacy is different between cultures.

Palen and Dourish (2003) points out that the nature of privacy as a dynamic process, regulating according to the current situation stands in contrast to how an information and communication technology system typically operates. The authors give examples like, file systems, email and cell phones and databases. These systems are built on fixed standards. The conditions in which they operate should not constantly be altered as if these systems are to successfully communicate with other systems. Standards and rules are often required to make things work. This discrepancy is important to keep in mind as contradicting requirements must be dealt with when designing systems that may involve privacy and identity management issues. This could also be a reason why privacy could be considered poorly managed in many technical systems of today related to personal information.

3.2 Classification

To be able to make experiments and support the thesis some technological concepts need to be defined. Classification algorithms and the concepts for technical analysis of text, often referred to as “Natural Language Processing”, NLP are explained here.

3.2.1 Support vector machines

As classifying data is a common task, there exist several methods doing classification and which method is most suitable depends on the kind of data that is to be classified. Support vector machines (SVM) are one method for classification. The idea behind SVM can be explained by imagine data represented in a space of dimension *n* where *n* is the number of attributes that is available in the data. The classifiers job is to find planes in that space that distinguish between the instances of the data the best possible way and make up the classes. It is a method that is useful especially when working with very high dimensional data. (Pang-Nin et. al, 2005)

Figure 1 shows the conceptual idea with support vector machines. Two classes are displayed, what is looked for is the plane (line in this 2-dimensional example) that maximizes the margin[1] to the classes. In this example we see that the solid line in the middle fulfils that. This is a very simple example but it illustrates the idea behind support vector machines. Data of much higher dimension, that cannot easily be visualized, it is not as intuitive as this example but the idea is the same.

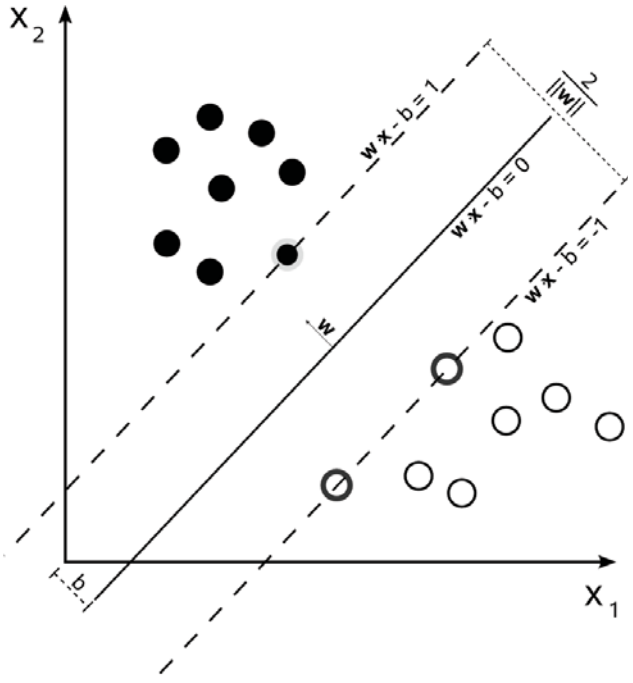

Figure 1 Maximum margin hyper plane (wikipedia.org)

To calculate the margins and maximize the distance different vector calculations is done. The simplest is to use linear model, kernel called. A non-linear can be used as well but it is more complicated and it is very easy to over fit the data. Therefore a linear kernel is often to prefer. (http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf).

3.2.2 Natural language processing

Natural language processing (NLP) is the comprehensive name for techniques used to technically analyze text. NLP analysis can be done by hand but the development of computer aided tools has made it achievable on a larger scale. What is done is that different linguistic measurements are calculated and comparisons are made. Examples of linguistic measurements are word length, sentence lengths, and type token ratio. Type token ratio is the relation of different words to the total number of words in a text. For example if an article consists of 1500 words where 500 of them are unique gives this a type token ratio of 500/1500 = 1/3 (Biber, 1987).

[1] $\frac{2}{\|W\|}$

4. Implementation

An implementation of a system enabling the user to be aware of her appearance on the World Wide Web would be of possible interest to a lot of people. What is presented in this chapter is a conceptual overview of such a system. What requirements are needed, what are the problems that must be handled in a good way.

4.1 Requirements

What is wanted from the system discussed in this chapter is to be able to have an open interaction with the pieces of information making up the identities. That means that a transparency in the flow of information within the system is crucial.

The result presented from the system is typically built from many different sources and is very complex. This feature of the results makes it desirable for the user to be able to interact with the output. The interaction could be to modify parts of the results, removing some pieces or putting focus on other pieces. This makes it necessary for the information going through the system to be traceable through the process. This requirement also makes the end user visualization an important aspect of the system.

4.2 Architectural overview

A system that could manage to fulfil the requirements transparency in the flow of information and traceable information thorough the process has to fulfil a number of subtasks. These tasks can be divided into four main functions. The tasks are presented in the order processing of the data is necessary to be done.

4.2.1 Data acquisition

In order to get information from the web a collection must be done. The data acquisition function collects web pages and presents the information selected through a first relevance filter. The relevance filter limits the data to information only related to the person in question, which is essential for the following functions. This limitation of the data makes the following steps of analysis accurate. This requirement is of great importance but it is hard to accomplish since for example there are many people with common names and to distinguish between them is a task that probably would need considerable attention.

4.2.2 Identity feature extraction

The web page itself is not a practical representation of the information that one want from the page. Therefore information pieces, features, must be extracted from the page. Identity feature function extracts the relevant information found on each web page from the data accusation function. The features gained from each web page by the identity feature extraction function are in a form that can be further analyzed and combined with features from other pages. This is done in the following steps of the process.

4.2.3 Creating compound facets of identities

The features extracted make the most sense in combination with each other. To form a useful foundation for analysis an aggregation of the features needs to be made by a function. The analysis could for example be categorizing keywords that are one of the

possible features. Since the features are traceable to the source web pages the equivalent combination of the web pages can be made. The way the features are combined, what pages are combined to a category, could make up a view of an individual's identity facets.

4.2.4 Representing the facets graphically

To make the gained information make sense to a user an intuitive visualization is necessary. This visualization is also crucial to make the other functions come to full need since it is of no good if the user can not comprehend the available information properly. The requirement of the system to have the abilities to have an open interaction with the information and the traceability of the data is also functions that must be handled in a good way by the user interface. This makes the function of graphical representation at least as important as the previous functions.

4.3 Case study

In order to practically find and identify representations of people on the web a case study were conducted. The case study evaluates the applicability of some methods that may be used as components of the system discussed in 4.2. The case study was done iteratively in three steps, using three setups. The outcome of each step makes the foundation for the following one.

4.3.1 Goal

The goal with the case study is to identify representations of people's facets of identities on web pages. This is done to test possible approaches to technically fulfil the functions of identity feature extraction and creating compound facets of identities. To evaluate the results are they compared to the expected result of a comparable naïve classification method. In this investigation were the results compared to a random classification with regard to the distribution of the data.

4.3.2 Delimitation

Investigating possible methods for all functions described in 4.2 beyond in the scope of this report. Therefore the "data acquisition" function and "Representing the facets graphically" are not discussed further. However, I don't argue that they are not important aspects of the system but for practical reasons delimitations must be made. What is focused on further is "identity feature extraction" and "Creating compound facets of identities". Possible methods to support and fulfil these functions are investigated in the case study.

4.3.3 Methodology

The purpose of the Identity feature extraction function is to find pieces of information on web pages that can be used to say something about an individual's identity. What possible pieces is that? And what aspects of an individual's identity are feasible to find from web pages? To approach these questions a Grounded theory approach is used. Grounded theory is a qualitative research method for the collecting and analysis of qualitative data. It is a bottom-up method with data as the starting point. By deeply analyzing the data, without any premade assumptions or ideas, structures and theories evolve. This makes grounded theory a typical deductive method (Charmaz, 2006). Grounded theory is popular for working with qualitative data when the researcher does

not know in advance how the data is constructed and to use it the best way. Grounded theory has had an increase in usage within research about information systems. (Jones, Hughes 2003)

4.3.4 Setup 1

The idea behind this thesis is to find representations of people in web pages. Naturally web pages are the data to work with. A collection of web pages as basis for analysis was created using the search engine Yahoo. Yahoo was chosen since it provides a useable API[2] for automating the retrieval. Google, for example, does not provide that. Personal names were chosen as search terms since that would increase the chance of the page retrieved actually to reflect a person. The names were chosen to higher the probability of finding web pages in English since pages in other languages were discarded (see Appendix for list of the names used). In this first setup 67 pages were retrieved. The pages were read individually while highlighting aspects that came to mind while reading them. This process was repeated several times to be able to make aggregated conclusions of the data.
Figure 2 shows a screen shoot of a web page that was analyzed in that way. Highlighted are keywords that give an idea of the content of the page. “Media consultant”, “marketing career”, “PR” are example of clues that indicates that this page is related to someone’s work and that it seems to be in the area of marketing and public relations.

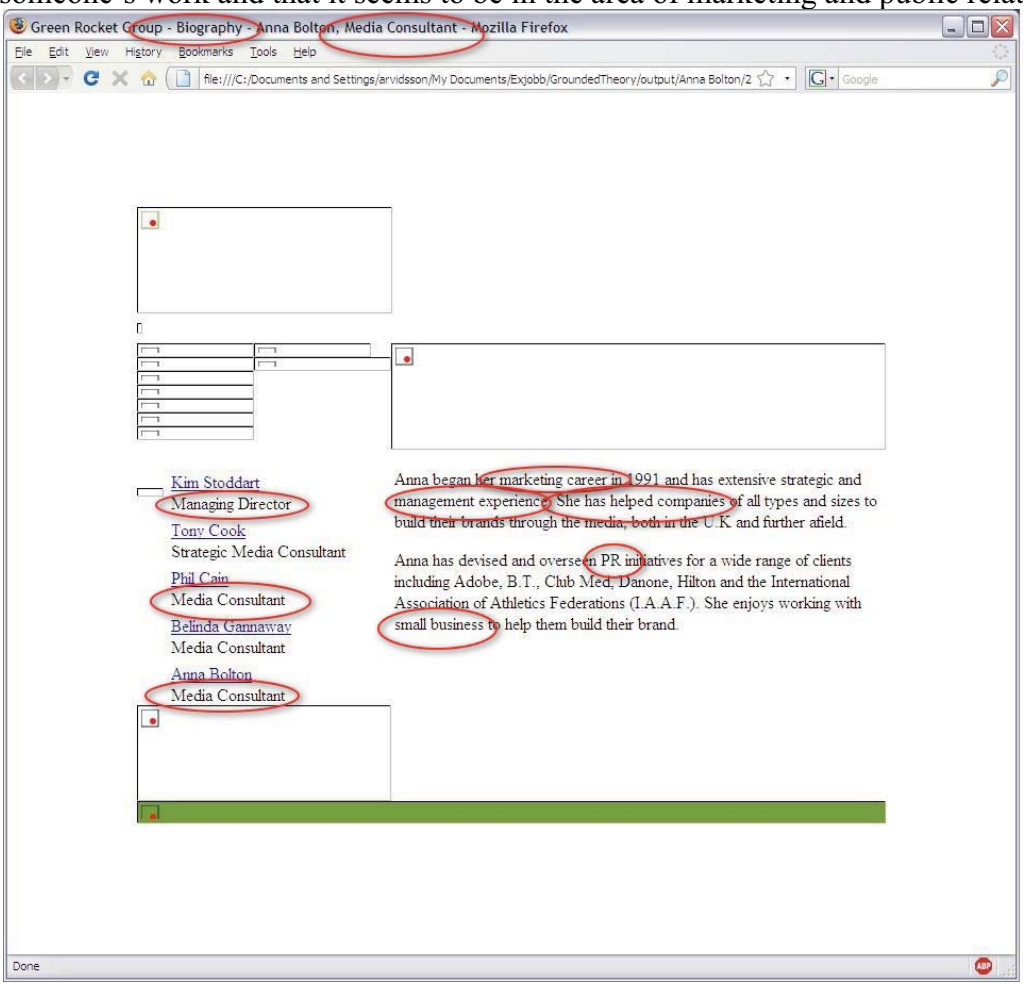

Figure 2 Web page with highlighted keywords.

[2] Application programming interface.

4.3.4.1 Result

The manual analysis of the pages revealed a lot of information. That was summarized into three dimensions of the pages: *sphere, categories* and *text style*.

4.3.4.1.1 Sphere

Sphere represents people's different personal environments, typically a person exists in a work or school related environment and an environment related to her family and friends. Beside these environments one also have information that is aimed at people that one doesn't know beforehand. These categories of sphere were named "Professional", "Personal" and "Public". When discovering keywords as, "title", "position", "employment", "customer", etc gives that an idea that the content of the page concerns a person work life and the page may be classified as "professional". Likewise where there pages with a lot of family names and family relationships as "sister", "father", "family", "friend" giving an idea of a personal page

4.3.4.1.2 Categories

A dimension of the pages that were discovered was topic categories. There were for example pages with information about literature and other pages contained information about football. By combining the topics found on pages related to a name one can get interesting information, for example if a person's interest in genealogy has had a greater impact on her web presence compared to her work in advertising that is not as visible.

4.3.4.1.3 Text style

Another aspect of the page is that the level of engagement, level of sentiment differs. In some situations, for example in a heated letter to the editor in the local newspaper, one writes very personal and with strong feelings. In a more formal text, for example a scientific paper the text is in much more neutral and objective. These differences should be possible to use as data for an analysis of a person's web sources.

4.3.4.1.4 Data set

To enable further analysis data sets of the results from setup 1 was created. The open source data mining software package Weka[3] was used for classifying the data set that was created by the first setup. Weka is a multi purpose machine-learning tool that is capable of using a number of algorithms. For the classifying tasks in this case study an implementation of support vector machines called libSVM[4] was used.

4.3.5 Setup 2

The results from the first setup gave the foundation for continued work. In the second setup Sphere was in focus. Sphere was selected as a first target since the other dimensions would require a lot more data collection. Topic categorization is also a fairly mature area (Ceci and Malerba 2007, e.g.).

What was needed to increase the chance of improving the classification was to extend the analyzable data, and increase the number of web pages to be able to make reliable analysis. This was done by a new analysis of a set of web pages. The pages were retrieved using the same method as in the first setup, but with new names introduced. A

[3] http://www.cs.waikato.ac.nz/~ml/weka/
[4] http://www.cs.iastate.edu/~yasser/wlsvm/

total of 250 pages were collected and analyzed. These pages were manually classified according to the sphere classes that were discovered in setup 1. Based on the perceived intended use each page was classified as personal, professional or public. These pages were merged into the data set created in Setup 1.

4.3.5.1 Result

To test the classification of sphere a test-set of 26 web pages manually classified the same way as the training set was tested against a training set of 226 pages. The libSVM classifier gave a result of 69.2 % correctly classified pages. The distribution of the training and the testing data is about 50 % of the pages are "Public". "Professional" and "Personal" stands for 25% each. A classification algorithm that randomly classifies the material to that proportions would statistically end up with 37,5% correctly classified classes.[5] See appendix for a log from the Weka run.

4.3.6 Setup 3

Setup 3 focused on approaches to increase the accuracy of the sphere categorization of setup 2. The text style dimension that came out from the first setup seemed to be interesting to investigate further. How does the text on the page actually appear? During this third iteration we re-analyzed the texts from setup 2. The pages were noted to provide very different experiences depending on its text. We decided to focus on the three different dimensions *simple text vs. complex text*, *non text vs. text* and *page type*. These dimensions were measured using natural language processing methods. Included counting numbers of words, type token ratio, average word length and average sentence length.

4.3.6.1 Simple text vs. complex text

As the name indicates this dimension distinguishes between pages with regard to the perceived difficulty of the text. An example of a text that seemed simple was a short biography. A complex text found was an arguing political article for example.

4.3.6.2 Non text vs. text

Despite its name do non-text pages contain text, but the text is not full sentences and paragraphs. Non-text pages can be pages with extensive lists or profiles with a certain layout that one typically not associate with text.

4.3.6.3 Page type

The web consists of many different kinds of pages. The page type dimension reflects this by categorizing pages in a number of categories found, for example:

- Media – pictures, videos.
- Listings&Directories – Lists of various kinds.
- Discussions – Forum posts, email exerts.
- Profiles – Personal profiles, LinkedIn, Facebook.
- Promotional - Advertising, selling products.
- Blogs – Blog posts.
- Articles – Newspapers, scientific.
- Genealogy – Family trees.
- Literary – Fiction like texts.

[5] 50%*50% + 2*(25%*25%) = 37,5%

- Biographies – Obituary, general biographies.

4.3.6.4 Result

Table 1 shows the results from text type classification. The number of pages analyzed was not as many in this setup due to the fact that not all pages could be analyzed or classified in the page type dimension.

Table 1 Setup 3 classification

	Simple text vs. complex text	Non-text vs. text	Page types
Correctly classified	64 %	93 %	41 %

”Simple text vs. complex text” and ”Non-text vs. text” is binary classes and the distributions of the classes are close to even. Therefore the baseline of a random classification is 50%. ”Page types” has 10 classes with a share for each of them varying from 2% to 27% making it harder to compare to a “straight” random selection. If the comparison to our classification is a naïve algorithm that classifies all instances as the largest class“ Profiles” that would give 27% correctly instances. See appendix for detailed results from the test with Weka logs.

5. Discussion

Finding representations of individuals on web pages is not a trivial task. What is investigated in the case study in this report are some possible ways for doing that. In the thesis of this report I claim that a combination of information extraction, data mining and natural language processing techniques are possible tools for finding these representations. What are shown are classifications of web pages with regard to sphere in Setup 2 and with regard to text styles in Setup 3. The results of these experiments are acceptable considering the amount of data that is analyzed and the comparison for each test. The comparisons of the classifications are done with awareness of the distribution of the data to make them better. Since all classifications are better than the comparisons the thesis is supported by the case study.

If one would combine the results of the classifications and make a comprehensive analysis the results that are presented in this report indicates interesting co-occurrences. For example, maybe more complex texts are common in combination with the professional sphere.

5.1 Issues with the investigation

The delimitation of the "Data acquisition" part in chapter 4.2 sidesteps one big problem that must be handled if one would implement the kind of system described. Several people can have the same name, how to distinguish between then? This does not influence my thesis, but it is an important aspect to keep in mind.
The possible amount of data is limited since all preparation and pre-classification of the data sets is made manually. With a larger data set the reliability of the case study could be improved since the thought of application must be able to cope with millions of web sources.

5.2 Ethical considerations

Aggregated personal information is very sensitive in it's nature. The main intended possible use of that kind of material in this report is a tool that may visualize the individual's appearance on the web. The primarily recipient is the individual in question herself, helping her understand how others perceive her from the information available on the web. It is of course not unimaginable that other people would be interested in this information as well. Recruiting situations is one example. Interest from other agents whose interest may not be in line with the individual are also a possibility. One can argue that this kind of issues questions if it is correct to conduct research that could be used in ways that may be to the individual's disadvantage. It is a fact that almost every technology could be used for destructive reasons. Nuclear power is an obvious example. However if no research that possibly could be used in harmful ways were conducted there would not be much research left. Therefore the potential issues must be kept in mind while evaluating and possibly implementing technologies of these kinds but the possible benefits are the motivation.

5.3 Future work

The analysis techniques used in the investigation solely depend on the text occurring on the web pages. The results would have been just the same if text files with the text

visible on the site were used instead. There is a lot of other information on web pages beside the text. Pictures, videos and layout are examples of this kind of information. A lot of digital documents contain Meta data, data about the data. Pictures often have information about the camera settings and the date. Another example is pdf-documents that may contain information about the author. All of these different sources of information may form a basis for classification. An investigation of data from a variety of sources and with a variety of types of data would make natural and interesting extension of this research. This would probably give interesting results, since one can imagine that a lot how human beings interpret information is by layout and pictures. The issue of distinguishing between persons with the same name maybe can be handled by face recognition image analysis for example. As discussed above this case study is done by a small amount of data. A bigger investigation with a substantially larger corpus of data, better reflecting actual circumstances on the web would be interesting to test the accuracy of the classifications on a larger scale and test the scalability of the methods described.

6. References

Altman, I. (1975), The Environment and Social Behavior. Brooks/Cole, Monterey, CA.

Altman, I. (1977). “Privacy Regulation: Culturally Universal or Culturally Specific?” Journal of Social Issues, 33 (3), 66-84.

Biber, D. (1987). Variation across speech and writing, Cambridge university press, New York.

Boyd, Danah. (2002). "Faceted Id/entity: Managing Representation in a Digital World." Cambridge, MA: MIT Master's Thesis. August 9, 2002.

Bylund, M and Karlgren, J and Olsson, F and Sanches, P and Arvidsson, C-H (2008) ”Mirroring Your Web Presence” In: CIKM 2008 Workshop on Search in Social Media (SSM 2008), 30 Oct 2008, Napa Valley, California.

Ceci, Malerba, (2007), Classifying web documents in a hierarchy of categories: a comprehensive study. Journal of Intelligent Information Systems, Volume 28, Number 1.

Charmaz, K. (2006). Constructing Grounded Theory: A Practical Guide through Qualitative Analysis, Sage Publications, London.

Jenkins, R. (2003). Social Identity 2nd ed. Rout ledge, London.

Hughes, J, Jones, S, 2003, Reflections on the use of Grounded Theory in Interpretive Information Systems Research. Electronic Journal of Information Systems Evaluation.

Palen, L. and Dourish, P. 2003. Unpacking "privacy" for a networked world. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '03. ACM Press, New York, NY, 129-136.

Pang-Ning, Steinbach, Kumar, (2005) *Introduction to Data Mining*, Addison-Wesley, Boston.

Figure 1 from Wikipedia.org, picture in public domain.
http://en.wikipedia.org/wiki/Image:Svm_max_sep_hyperplane_with_margin.png

Homepage of data mining package WEKA 2007-10-29
http://www.cs.waikato.ac.nz/ml/weka/

Implementation of LibSVM for WEKA 2008-04-03
http://www.cs.iastate.edu/~yasser/wlsvm/

Practical guide to Support Vector Classification
http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf 2008-04-03

Appendix

Names used for web retrieval

These names where used as search terms for acquiring the web pages:
"steven hughes"
"kyle spencer"
"brittany carlson"
"john moore"
"louise watson"
"richard keenan"
"carla lindsay"
"adam jackson"
"alexandra marley"
"peter charles"
"michelle brunton"
"benjamin lockhart"
"david wells"
"camilla johnson"
"bertha ray"
"maria simpson"
"mark abbott"
"katherine dodds"
"nathan riley"
"emma closs"
"martha isaac"
"joel rogers"
"carl peterson"
"pedro sanches"

WEKA Logs

Setup 2 Sphere

```
=== Run information ===

Scheme:       weka.classifiers.functions.LibSVM -S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M
40.0 -C 1.0 -E 0.0010 -P 0.1 -Z
Relation:     C:\Documents and Settings\arvidsson\My
Documents\Exjobb\Tests\Latest\training_nofilter_noweight.libsvm-
weka.filters.unsupervised.attribute.NumericToNominal-Rlast
Instances:    226
Attributes:   26050
              [list of attributes omitted]
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 5.52 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          18               69.2308 %
Incorrectly Classified Instances         8               30.7692 %
Kappa statistic                          0.4721
Mean absolute error                      0.205
Root mean squared error                  0.4529
Relative absolute error                 49.4334 %
Root relative squared error             96.5615 %
Total Number of Instances               26

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  1         0.571      0.6       1         0.75       0.714    0
  0.429     0          1         0.429     0.6        0.714    1
  0.429     0          1         0.429     0.6        0.714    2

=== Confusion Matrix ===

 a  b  c   <-- classified as
 12  0  0 |  a = 0  Public
  4  3  0 |  b = 1  Professional
  4  0  3 |  c = 2  Personal
```

Setup 3 Text vs. non-text

```
=== Run information ===

Scheme:       weka.classifiers.functions.LibSVM -S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M
40.0 -C 1.0 -E 0.0010 -P 0.1 -Z
Relation:     Linguistic Features-weka.filters.supervised.attribute.AttributeSelection-
Eweka.attributeSelection.CfsSubsetEval-Sweka.attributeSelection.BestFirst -D 1 -N 5
Instances:    27
Attributes:   5
              numWords
              averageSentenceLength
              percentageStopWords
              daleChall
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          25               92.5926 %
Incorrectly Classified Instances         2                7.4074 %
Kappa statistic                          0.8516
Mean absolute error                      0.0741
Root mean squared error                  0.2722
Relative absolute error                 14.6893 %
Root relative squared error             53.9134 %
Total Number of Instances               27

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.929     0.077      0.929     0.929     0.929      0.926    Text
  0.923     0.071      0.923     0.923     0.923      0.926    Non-text

=== Confusion Matrix ===

  a  b   <-- classified as
 13  1 |  a = Text
  1 12 |  b = Non-text
```

Setup 3 Simple vs. complex

```
=== Run information ===

Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
-K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
Relation:     Linguistic Features-weka.filters.supervised.attribute.AttributeSelection-
Eweka.attributeSelection.CfsSubsetEval-Sweka.attributeSelection.BestFirst -D 1 -N 5
Instances:    14
Attributes:   2
              colemanLiau
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

SMO
Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: Simple, Complex
BinarySMO

Machine linear: showing attribute weights, not support vectors.

        -1.628  * (normalized) colemanLiau
 +       0.5315

Number of kernel evaluations: 27 (35.714% cached)

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           9               64.2857 %
Incorrectly Classified Instances         5               35.7143 %
Kappa statistic                          0.186
Mean absolute error                      0.3571
Root mean squared error                  0.5976
Relative absolute error                 70.7547 %
Root relative squared error            117.332  %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  1         0.833      0.615     1         0.762      0.583    Simple
  0.167     0          1         0.167     0.286      0.583    Complex

=== Confusion Matrix ===

 a b   <-- classified as
 8 0 | a = Simple
 5 1 | b = Complex
```

Setup 3 Page type

```
=== Run information ===
Scheme:       weka.classifiers.functions.LibSVM -S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M
40.0 -C 1.0 -E 0.0010 -P 0.1 -Z
Relation:     Linguistic Features
Instances:    120
Attributes:   22
              numWords
              averageWordLength
              longWordsCount
              longWordsPercentage
              averageWordSyllables
              highSyllableWordsCount
              highSyllableWordsPercentage
              averageSentenceLength
              typeTokenRatio
              numStopWords
              percentageStopWords
              numFamiliarWords
              percentageFamiliarWords
              numDigits
              fogIndex
              daleChall
              daleIndex
              fleschKincaidGrade
              fleschReadingEase
              colemanLiau
              bormuthGrad
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 0.34 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          49               40.8333 %
Incorrectly Classified Instances        71               59.1667 %
Kappa statistic                          0.2338
Mean absolute error                      0.1183
Root mean squared error                  0.344
Relative absolute error                 70.7721 %
Root relative squared error            119.2704 %
Total Number of Instances              120

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0          0          0          0          0          0.5      Media
  0          0.009      0          0          0          0.495    Listings&Directories
  0          0          0          0          0          0.5      Discussions
  0.813      0.443      0.4        0.813      0.536      0.685    Profiles
  0          0          0          0          0          0.5      Promotional
  0          0          0          0          0          0.5      Blogs
  0.529      0.107      0.45       0.529      0.486      0.711    Articles
  0          0          0          0          0          0.5      Genealogy
  0          0          0          0          0          0.5      Literary
  0.56       0.211      0.412      0.56       0.475      0.675    Biographies

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j   <-- classified as
  0  0  0  5  0  0  0  0  0  1 |  a = Media
  0  0  0  8  0  0  0  0  0  2 |  b = Listings&Directories
  0  0  0  1  0  0  0  0  0  1 |  c = Discussions
  0  1  0 26  0  0  2  0  0  3 |  d = Profiles
  0  0  0 11  0  0  1  0  0  7 |  e = Promotional
  0  0  0  1  0  0  2  0  0  0 |  f = Blogs
  0  0  0  3  0  0  9  0  0  5 |  g = Articles
  0  0  0  3  0  0  0  0  0  0 |  h = Genealogy
  0  0  0  0  0  0  2  0  0  1 |  i = Literary
  0  0  0  7  0  0  4  0  0 14 |  j = Biographies
```