Automatic Analysis and Visualization of Stylistic Genres

Erik Johannesson & Christopher Wallström
{erik, cw}@ viktoria.informatik.gu.se

Abstract

In this paper we are going to present a method capable of visualizing and automatically determining similarity between stylistic genres in Swedish and English text documents. The method has also been implemented and tested with good results. It is based on linguistic methods and uses only the raw word content of documents, meaning that it requires no meta-tags or other features set by the authors of documents.

Keywords: IRIS, genre, information retrieval, information visualization
BRT Keywords: AB, AD, AE, AI, CB, DB, HB, IB.

Introduction

The realm of information has turned into one of chaos where the only real efforts to bring some kind of order are the search engines that roam the internet, collecting information guided by meta-tags and word content. This approach to information extraction encourages misuse of meta-tags and abuse of frequently used keywords. What is needed is a new way of viewing documents, not as a string of keywords but as something greater than the sum of their parts. Documents contain more information than is readily visible, syntactic and semantic properties are the kinds of abstract information that can tell us a lot about documents if we use them wisely. One way of using this kind of underlying abstract information in documents is to develop new methods that incorporates linguistic theories that helps us capture this meta-data. This is the approach we have chosen.

We have developed a linguistic method that from the raw word content of a document extract enough stylistic information to be helpful both in classification, comparison and visualization of genres. The method is also easy to modify, expand and to translate into different languages. The primary language for the method is English, but we have also translated it to Swedish, and tested it with favorable results.

The abstract information we wanted to extract from documents was the information that denotes stylistic genre. The reason stylistic genre seemed interesting was that we felt a need for a tool that could distinguish between genres such as research papers, informal letters, news articles etc. We also needed a method to compare different genres to get an understanding of what each genre denotes, preferably in a quantitative way, and also determine the degree of similarity between genres.

The aim of this paper is to present and describe how we have developed an automatic genre evaluation and classification system. It also presents results from tests conducted at a small sample of documents, collected from the internet, to give significance to the method we have developed. Its hidden agenda is to illuminate the advantages and possibilities of using computational linguistic methods for information retrieval and visualization. The audience this paper is intended for is people who are working in the field of information retrieval or visualization, and are interested in new ways of extracting relevant information from text documents.

In this paper we will begin by addressing some of the different opinions of what genre really is, and define what we mean when we say “stylistic genre”. Then we will present what kind of information we extract from documents to make the stylistic comparisons, and why this information seems important. Next we present two different analyzes of the data. Finally we present conclusions and thoughts about future possibilities of the method we have designed.

Related work

Information retrieval and text filtering has been thoroughly examined and studied since the beginning of the computer society. We will here give a short presentation that positions our work in the field.

Douglas W. Oard (1997) gives a broad and rather comprehensive overview of the field of information retrieval and text filtering. He develops a conceptual framework for text filtering practice and research, and reviews present practice in the field. The method of analyzing genre developed here can be used for information filtering, information retrieval or exploration, as based on the classifications in Oard (1997:6).

Douglas Biber (1986, 1988) makes a serious effort to find stylistic variations and correlations in different English texts, both spoken and written. He uses hand tagged material and factor analysis to get these results. The method described in this paper uses a subset of the linguistic features Biber uses, and as a result it can be used on non-hand tagged material.

Jussi Karlgren et al (1998) developed a program using the same linguistic theories as we. The difference between our approaches is that they assign each document to one exclusive genre instead of making comparisons between genres and using similarity measures. Another difference is that Karlgren et al (1998) only developed the method for English texts. Brett Kessler et al (1997) has also developed a similar method but it is also limited to English documents and sorts the documents into only one genre, although in a slightly different way.

Which kind of genres do we explore?

After determining which kind of genres we thought would be useful we sought to find a definition that corresponds to our view of the term genre. Carolyn Miller (1994) writes that the urge to classify is fundamental and that classification is necessary to language and learning. She furthermore argues that a theoretically sound definition of genre must be centred not on the substance or the form of discourse but on the action it is used to accomplish. Carolyn Miller proposes that in rhetoric the term ‘genre’ should be limited to classification based in rhetorical practice and therefore it has to be an open classification system rather than closed and organized around situated actions. This correlates to our pragmatic view of investigating documents to analyze genre.

Miller (1994) defines genre through recurrent rhetorical situations and points out that it is necessary to reject materialistic tendencies in situational theory. Stebbins (1967:154) notes that “objective situations are unique”, which means that these situations cannot recur, a typically materialistic point of view. Our understanding of situations as somehow ‘comparable’, similar, or ‘analogous’ to other situations imply recurrence, Miller (1994) notes. According to her, recurrence is an intersubjective phenomenon, a social occurrence that cannot be understood on materialistic terms. She also says “situations are social constructs that are the result, not of ‘perception’, but of ‘definition’”.

In order to interact and respond to new situations, we need to have undergone certain typification processes. Miller (1994) states that these typification processes are what make us able to create recurrence, analogies and similarities. The representations of the types are semiotic structures that help us generate our response to any number of infinite possible situations. Shared types are what make communication through language possible.

Berkenkotter and Huckin (1995) widen the definition of genre and present a framework for elaborate genre studies. This framework consists of five parts: Dynamism, Situatedness, Form and Content, Duality of Structure and Community Ownership.

- Dynamism: The aspect of dynamism is summarized by Berkenkotter and Huckin (1995):

  > Genres are dynamic rhetorical forms that are developed from actors' responses to recurrent situations and which serve to stabilize experience and give it coherence and meaning. Genres change over time in response to their users' sociocognitive needs (p. 4).

  This implies that genre is an open class constantly changing and evolving, with new genres emerging and old ones decaying. Despite this Berkenkotter and Huckin (1995) argues that genres must be stable enough to capture those aspects of situations that tend to recur.
- Situatedness Genre knowledge is a form of situated cognition that according to Berkenkotter and Huckin (1995) is a product of the activity and situations in which it is produced. Bakhtin (1986) distinguishes between primary and secondary speech genres. The primary genres are those which we use in daily communicative activities and are dependent on the context in which they are produced. Secondary genres, on the other hand, are those that are excluded from the situated context in which the primary ones exist. Through this exclusion they gain the ability to function both irrespective of the context in which they are produced and over a longer period of time (e.g. scientific articles and written forms of organizational communications).
- Form and Content A genre encompasses both form and content. The content that is appropriate for a genre changes over time, perhaps even more than form changes. Furthermore genre knowledge encompasses not only knowledge of formal conventions but also knowledge of appropriate topics and relevant details as well. Matters of content-epistemology, background knowledge, surprise value, timing (kairos)-all influences the selection and use of formal features in the instantiation of particular genres.
- Duality of Structure Berkenkotter and Huckin (1995) adapts Giddens (1979, 1984) view that duality of structure is at the center of structuration theory in sociology. Giddens (1979) argues that: “Structure is both medium and outcome of the reproduction of practices. Structure enters simultaneously into the constitution of the agent and social practices, and ‘exists’ in the generating moments of this constitution” (p. 5). In other words when we act in the context of an organization or an institution we adapt certain social structures and at the same time reproduce these structures in our own behaviour thus shaping them through our own reality filter. Genre emerges out of practice and shapes at the same time that practice (Ljungberg 1997).
- Community Ownership Genre use in the perspective of whom is using them reveals a great deal about a discourse community's norms, epistemology, ideology, and social ontology (Berkenkotter Huckin 1995). Ljungberg (1997) argues that understanding genre is closely related to understanding the community “owning” it.

By using this framework as a base to study the genres we are interested in we hope to understand what type of text documents that denote a certain genre. Since we are analysing text documents we can not encompass all these aspects of genre analysis but we hope to see traces of those aspects undetectable by our method.

Even though we see genre as an open class constantly changing and evolving, with new genres emerging and old ones decaying, we believe that it is meaningful to analyze and define genres, as they are fairly static in time. Changes in different aspects of our society can also be studied through continuous analysis of genres.

How did we extract the stylistic information from documents

The primary question we had to answer was what kind of abstract information can be extracted from a document to help us determine its stylistic genre? Previous research by Douglas Biber (1988) shows that certain linguistic features are clearly connected to what we perceive as stylistic genres. Biber (1988) identified six underlying textual dimensions in English texts through factor analysis of sixty-seven linguistic features in over 450 different texts of approximately 2 000 words each. Douglas Biber used hand-tagged material and of the sixty-seven original features twenty-nine do not require hand-tagged texts or automatic syntactic or semantic analysis. Since the aim of our work was to create automatic genre analysis of texts as an aid for information retrieval we could not assume that the texts we were to analyze would be tagged regarding to syntax and semantics. Due to practical limitations such as the time constraints of real time user interface and also the limited time for developing the programs we chose only to implement the twenty-nine linguistic features that are relatively straightforward. Typical features are:

> place adverbials, time adverbials, first person pronouns, second person pronouns, possibility modals, necessity modals, predictive modals, public verbs, private verbs, suasive verbs, seem/appear, synthetic negation, analytic negation, type/token ratio, and word length.

In order to be able to analyze Swedish texts we could not just translate the English words associated with each feature, but had to do our own deep study of the Swedish language to find the words that correspond to each of the features. The first step was to study “A comprehensive grammar of the English language” by Quirk et al. (1985), which Biber used to find the linguistic features, in order to get a first person perspective of what the features really meant. Next we used our common sense, a lot of grammatical books and some internet-based resources to find the Swedish words that denotes the linguistic features. One potential problem that we did not have to deal with was if one of the linguistic features would not have a comparable Swedish counterpart. Swedish and English are rather similar languages, at least in this aspect.

Biber uses the separate features to identify six textual dimensions ‘each of which defines a different set of relations among texts’ (Biber 1988:169). He further notes that ‘These dimensions do not represent all of the differences defined by the original 67 linguistic features. Rather the dimensions are abstractions, describing the underlying parameters of variation in relatively global terms.’ Our goal on the other hand was to tell genres apart automatically, and thus we wanted as exact descriptions of the genres as possible. This means that taking the more abstract view of the textual dimensions only would cause us to lose some focus in the comparison.

The next step was implementing all of this to make the extraction of the stylistic information automatic. We chose to use Perl, which is a programming language that fulfills the criteria we have set. This means that it should be easy to use, easy to understand and modify, capable of handling large amounts of text, fast, easy to download documents from the internet with and easy to make programs that can be executed from a web browser in. Both Java and C++ fulfil these criteria but we felt that the amount of work needed to implement it in these languages was much higher. We implemented a program that first decided what language the text was written in, via a simple algorithm. Then it counted the number of times the words from each feature occurred and normalized it to an average length of 1000 words in order to get comparable results from both short and long texts. Next it calculated the average word length and the type/token ratio. This resulted in twenty-nine real numbers that represents the stylistic features of the document.

Evaluation of the method used for extracting genre

The goal of our research was constructing an automatic method to analyze the genre of text documents. The method would also have to be robust to be useful. To prove these abilities in our theoretical model we had to make empirical research. First we will present the evaluation methods we used and then the results of the evaluations. We chose to use two different and unrelated methods for the analysis and visualization of the data our method generates. The first method uses linear algebra to compare two of the calculated vectors to each other. The second method uses Kohonen (1995) self-organizing maps to visualize how the various documents correlate to each other.

The comparison method

The basis for this method is that sometimes you want to know how well a certain document correlates to a predefined genre. This is could prove to be very useful if you want to find a certain document e.g. a research paper concerning dog anatomy and do not want the search engine to display documents advertising dog merchandise or announcing dog exhibitions. Then you could use this feature as a filter only allowing documents of the genre ‘research paper’ to be presented.

In our test we used three genres as a base for the comparison: news articles, research papers and FAQ:s. We chose these genres mainly because they are widely used on the internet and can be clearly distinguished (Dewe 1998). The FAQ genre is also interesting since it is a genre that has emerged mainly on the internet.

The first step to complete this study was to collect the documents that would serve as templates for the comparison. A rather subjective approach was used to solve this problem: We collected five documents in each language that we thought was representative for each genre. In this study we wanted to put the emphasis on the usability of the method, we did not try to gain knowledge about the true meaning of what denotes a certain genre. For each of the three genres we calculated a template vector that was the mean vector of the five documents denoting each genre for both Swedish and English.

The second step was to find ten documents (not those we used for the template vectors) of each genre for each language to compare with the three genre template vectors. The documents was carefully selected from various sites on the internet and we only used documents that we really thought was representative for each genre.

The third step was to compare the template vector of each genre with the vectors of the documents we want to compare. We used cosine of the angle, a linear algebraic method, to compare the vectors. All elements of the vector are positive and therefore lie within the first quadrant and consequently this comparison results in a real number ranging from 0 to 1, where 1 is a perfect match (0 degree angle) and 0 is the worst match (90 degree angle). Each document was compared with the template vector for each of the genres within the same language.

Kohonen self organizing maps

The goal with the Kohonen self-organizing map analysis (Kohonen 1995) was to see if there was any relevant information hidden in the data we extracted from documents. This analysis seemed interesting because the self-organizing properties of the map makes the results unrelated to any preconceptions of the documents that are to be sorted in. The Kohonen self-organizing map does not need any predefined genres or groups of the documents but sorts them itself, so the result is completely unbiased. How the map sorts the documents is directly related to what the abstract information that is extracted from the documents really denote.

A Kohonen self-organizing map is a two-dimensional neural net consisting of a number of nodes linked in a two-dimensional structure much like a regular map or a fishing-net. Each of the nodes has a template vector of the same dimensionality as the vectors the map is supposed to organize. To make a map the following steps are used:

- Each node in the neural net is initialised either with a random value for each dimension or in some pattern (e.g. incrementally from left to right on the map).
- A randomly selected document vector, the input vector, is compared to each node on the map in some fashion, e.g. euclidian.
- The node with the most similar vector, the winning node, and all nodes that lie within a predetermined learning radius are changed based on a learning rate to become more similar to the input vector.
- Step three and four is repeated a predetermined amount of times while continually decreasing the learning radius and the learning rate.

Early on, the map gets its overall structure while the learning rate is high and the learning radius is big, and while the learning radius and the learning rate decreases the map gets its more specific local character.

When the Kohonen map is finished it sorts every document into the node that is most similar to its stylistic vector. When all of the documents are sorted into the map it also displays different shades of gray on the spaces between the nodes. The shade depends on how large the difference is between the nodes, the darker the greater the difference.

Results of the comparison method

Here are the results of the tests we conducted presented. The results are ordered according to which template vector that was used. First we present the results of the comparison to the template vector for Swedish and English news articles, then for research papers and last for the FAQ:s.

The English news articles (table1 and figure1) were clearly distinguished in comparison to the FAQ:s, while the difference between news articles and research papers was smaller. This smaller difference can perhaps depend on that both news articles and research papers are clearly informative in their disposition and actively tries to integrate information in the text. This in opposition to FAQ:s which present the information necessary to answer the question in a more direct way, which is one of the differences the method was designed to handle.

The difference between the Swedish news articles and the FAQ:s was lower than the difference between the English equivalents. This difference can partly depend on that the sample of Swedish FAQ:s is not very large. Perhaps the FAQ:s that has been dominant in this test is not representative for the Swedish FAQ:s or the Swedish FAQ:s are more closely related to news articles than the English ones. The values do however display a significant difference between news articles and FAQ:s and that difference is more important than the comparison of the differences between Swedish and English that we know exists. Below follows diagrams and tables to visualize the differences (figure 1-2 and table 1-2):

Figure 1

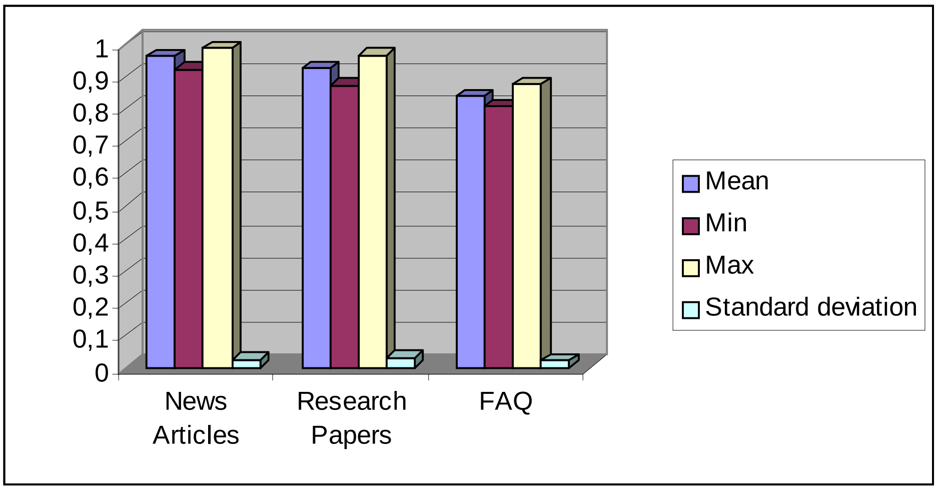


Table 1

Eng News		Mean	Min	Max	Standard deviation
	News Articles	0,960155	0,918635	0,985904	0,023032
	Research Papers	0,923389	0,871164	0,963859	0,03043
	FAQ	0,837078	0,805723	0,872625	0,021038

Figure 2

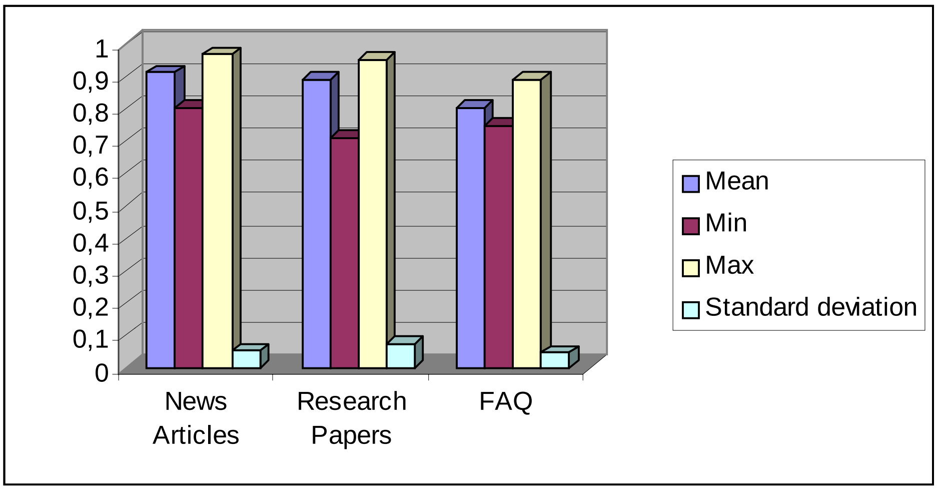


Table 2

Swe News		Mean	Min	Max	Standard deviation
	News Articles	0,911757	0,803054	0,965304	0,052938
	Research Papers	0,887979	0,711498	0,949668	0,073829
	FAQ	0,803858	0,747638	0,890683	0,045244

The difference between research papers and news articles was not very big (table 3-4 and figure 3-4). The reason for this is probably the same as those discussed above and that the structure of both research papers and news articles are similar. There is also the aspect that all documents we have retrieved are published for the same medium, the internet. In the test conducted it showed that research papers and news articles sometimes overlapped. The comparison of the template vector for research papers with research papers got a lower mean result than the comparison between the news article template vector and news articles. Th is gives more indications of research papers and news articles being stylistically similar. The mean value in both tests (both news articles and research papers) for both Swedish and English was higher for the template vector that represented each genre however. Below the values are presented in both diagrams and tables (figure 3-4 and table 3-4):

Figure 3

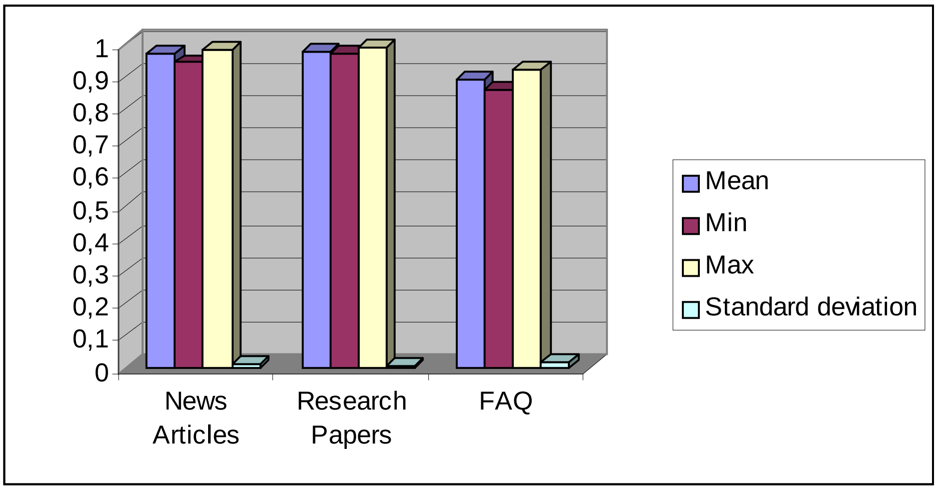


Table 3

Eng Research		Mean	Min	Max	Standard deviation
	News Articles	0,968221	0,941109	0,981387	0,012484
	Research Papers	0,977381	0,968975	0,986278	0,005979
	FAQ	0,887666	0,85782	0,919834	0,020064

Figure 4

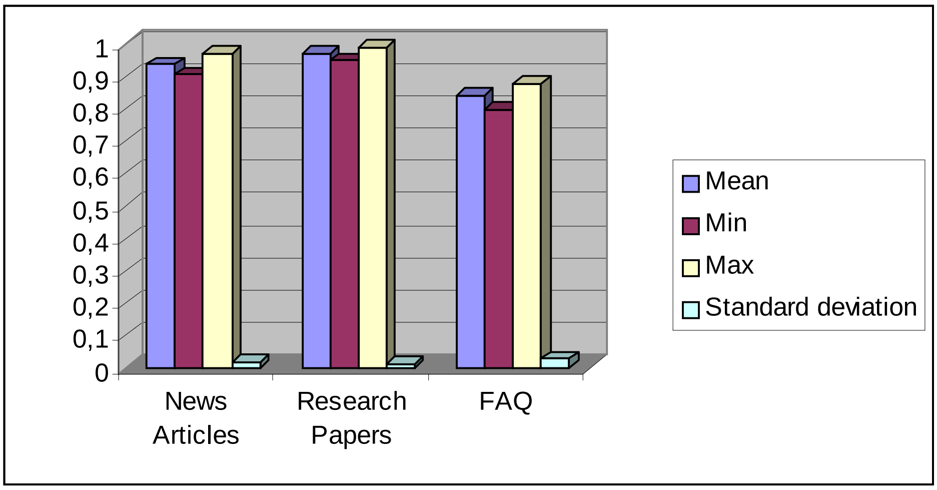


Table 4

Swe Research		Mean	Min	Max	Standard deviation
	News Articles	0,935809	0,905082	0,968179	0,018111
	Research Papers	0,968669	0,948364	0,986669	0,013339
	FAQ	0,838693	0,798054	0,875904	0,027247

As the two earlier tests has shown the FAQ:s genre is somewhat different from both news articles and research papers. The English FAQ:s (table 5) did result in a higher mean value than the Swedish ones (table 6) in comparison with the template vector. The reason for this can be that the number of sample FAQ:s available is by far much greater for English than Swedish, as discussed before. It can also be that the genre that denotes a FAQ is different for English than it is for Swedish and not as homogenous. The difference between FAQ:s and the other sample documents is equal for both Swedish and English (figure 5-6 and table 5-6).

Figure 5

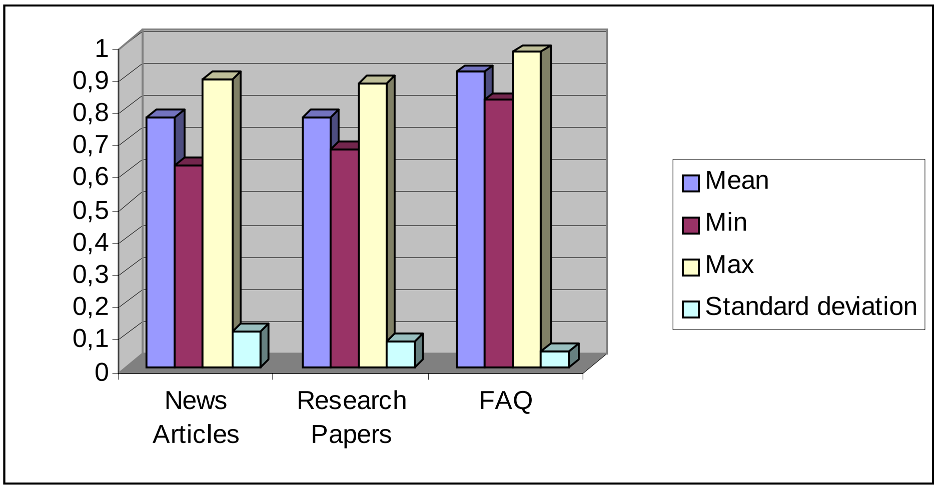


Table 5

Eng FAQ		Mean	Min	Max	Standard deviation
	News Articles	0,77019	0,622683	0,890747	0,110938
	Research Papers	0,772534	0,672931	0,87509	0,078566
	FAQ	0,911026	0,823786	0,972172	0,0477

Figure 6

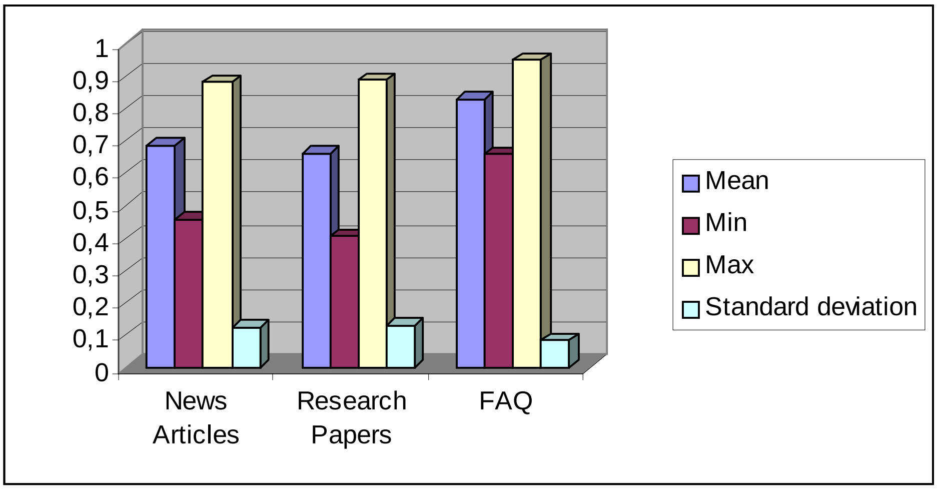


Table 6

Sve FAQ		Mean	Min	Max	Standard deviation
	News Articles	0,683578	0,455444	0,880119	0,12292
	Research Papers	0,660252	0,405274	0,885485	0,130207
	FAQ	0,826393	0,656931	0,948818	0,082589

Conclusions from the results of the comparison method

Through this test the sample documents has proved to correspond with the highest value to the right sample vector. Both news articles and research papers are very similar however. This similarity can probably be explained by how they are used. Both news articles and research papers are published on the internet and are most likely conformed in some way to make the best use of this medium. If we should analyze them in their usual medium (newspapers and scientific publications) we could find them to diverge more from each other. FAQ:s has proven to be easily distinguishable from the other to genres we analyzed. This is most likely a result from both the fact that their disposition is very different (questions and answers in a long row) and that FAQ:s remain in their original form as they were made for the internet from the beginning and not adapted to suit it. It is therefore not easy to compare this genre with two so similar genres as news articles and research papers. To sum it up we have come to the conclusion that all tests conducted with this comparison method have been successful and that the method proves sound and reliable.

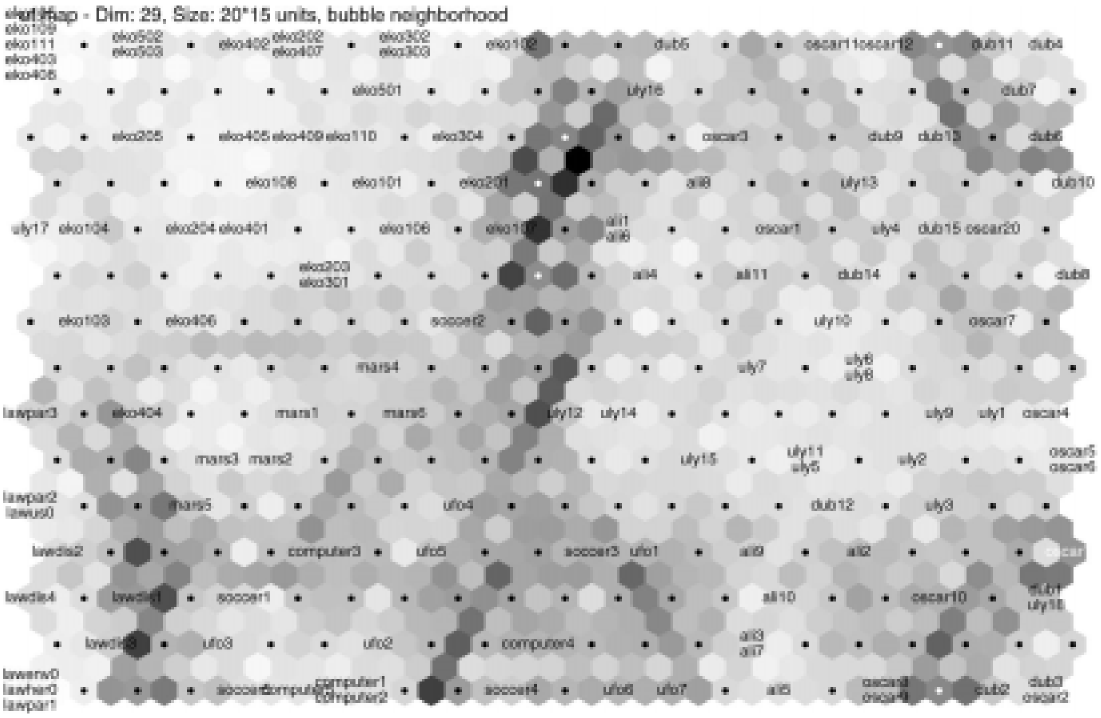
Results of Kohonen self organizing maps

The documents that we used for analysis with Kohonen self-organizing maps were first processed by our program. This resulted in a 29 dimensional vector that represent the stylistic form of each document. The value of each dimension of the vector corresponds to the amount of occurrences per 1000 words in the document of one of the 29 linguistic features we use. The documents were taken from 10 diverse categories, on the map they are named as follows:

- ekoXY where Y is a chapter from book number X in “Wealth of Nations” by Adam Smith.
- soccer are various documents from the internet about soccer.
- marsX where X is a chapter from Percival Lowell’s book about the planet Mars.
- ulyX where X is a chapter from James Joyce’s book “Ulysseus”.
- dubX where X is a chapter from James Joyce’s book “Dubliners”.
- aliX where X is a chapter from Lewis Carroll’s book “Alice in Wonderland”.
- oscar are various novels by Oscar Wilde.
- ufo are various documents form the internet about flying saucers etc.
- computer are various documents from the internet about computers.
- lawyyyX where X is a legal text about yyy which is environment, disabled, heritage and parks.

A total of 122 documents were collected from these categories, and the vectors calculated from these documents where processed through Kohonens program for generating self-organizing maps. We chose a map size of 20*15 hexagon nodes and the program generated the following map (figure 7):

Figure 7



On this map that resulted from our analysis it is clearly visible that fiction gathered on the right side, economic texts gathered on the top left part, mars, ufo, computer and sports related articles on the lower left part, and the law texts on the lower part of the left rim. The individual types of texts have clustered very well and the border between fiction and the other types of text is very sharp.

To get a better overview of this map we have divided it into four equally sized parts (see figure 8):

ali	0	ali	5
computer	0	computer	0
dub	0	dub	11
eko	31	eko	0
law	0	law	0
mars	0	mars	0
oscar	0	oscar	6
soccer	1	soccer	0
ufo	0	ufo	0
uly	1	uly	4

ali	0	ali	6
computer	5	computer	0
dub	0	dub	4
eko	1	eko	0
law	10	law	0
mars	6	mars	0
oscar	0	oscar	8
soccer	3	soccer	1
ufo	4	ufo	3
uly	0	uly	13

On this image you can clearly distinguish that the economic texts dominate the upper left quadrant, the fiction texts are all, but one, located in the two right quadrants and all but a few articles and law documents is located in the lower left quadrant. These results speak for themselves.

Conclusions from the results of Kohonen self organizing maps

We think that this map shows very well that the information that denotes the genre of a text can be found among the 29 features that we use. Using neural nets to find patterns among documents combined with the 29 features seems to be very successful and is something that very well could be used to find genres in many different environments. This method seems outstanding in distinguishing between these types of documents.

Conclusions and future work

New ways of information retrieval have to be based on and use the kind of classification methods that are natural to humans. Sorting documents according to genre is a natural process that should be supported by the programs that help you retrieve information.

Because of this we have chosen to create a method and program that based on linguistic theories manages to discern which genre or genres a document fits within or resembles most by using the method described above.

Words have both a meaning and a function in language. The method we have developed uses the functional aspect of words in written discourse. Miller (1994) supports the idea that genre can be distinguished by determining which functions certain words have. By applying the results of Biber (1988) on the functions of words we have implemented a method that both visualizes and compares genres. Since genre is recurrent rhetorical action we assume that documents using words with similar functions aim to perform similar actions and therefore belong to the same genre.

This paper has described a way of comparing documents based on their stylistic content, as well as a way of visualizing the stylistic content of a document. Our program is designed with an open solution to make it easily adaptable to most applications where genre analysis is useful.

Two examples of applications where we are using our genre analysis program are:

- Genre visualization for internet/intranet documents:
  Here the program helps by showing what is 'behind' a link before you click on it. The program could also be used to sort or filter links according to genre.
- Visualizing large document collections:
  The program uses either a neural net or clustering algorithms to automatically organize documents based on the similarity in genre. This is very useful to get a good overview of what types of documents that exist for example on an intranet.

We are at the moment involved in a project with the purpose to analyze ad hoc genres on the intranet of a large Swedish corporation. By adapting our program to work properly on their intranet we hope to discover which genres they use and how they are using them. We will not try to categorize the documents based on some preconceived notion of what genres might exist but rather allow them to form their own genres in this part of the organizational community.

We believe that new techniques like the one we have used is necessary to help us navigate the constantly growing sea of information. More and more people are realizing that the search engines of today are not sufficient for their needs. People start to rely on portals where the information is classified into given categories. The problem with portals is that you are dependent on the information and classification that the portal administrator chooses to publish. This can be influenced by commercial interests and may then not at all be suitable for your needs. If this classification process can be made automatic the human workload will decrease and the amount of documents processed will greatly increase and objectivity will in some sense also be achieved. We predict that more and more of these more intelligent techniques will be implemented and have an impact on information retrieval as we know it.

References

Bakhtin, M. M. 1986 (Originally published in 1952). *Speech genres and other late essays.* Austin: University of Texas Press.

Berkenkotter, C. & Huckin, T. (1995*). Genre knowledge in disciplinary communication: Cognition/culture/power*. Hillsdale, NJ: Erlbaum.

Biber, D. 1986. *Spoken and written textual dimensions in English: Resolving the contradictory findings.* Language.62. 384-414.

Biber, D. 1988. *Variation across speech and writing.* New York: Cambridge University Press.

Dewe, J. (1998-10-05) *En prototyp för att klassificera dokument från WWW med avseende på genre och ämne*, http://www.student.nada.kth.se/~d92-jde/examensarbete/DropJaw.html

Douglas W. Oard. 1997 *The state of the art in text filtering. User Modeling and User-Adapted Interaction.* University of Maryland, College Park, MD, U.S.A.

Giddens, A. 1979. *Central Problems in Social Theory*. London: Macmillan.

Giddens, A. 1984. *The Constitution of Society*. Cambridge: Polity Press.

Karlgren Jussi, Bretan Ivan, Dewe Johan, Hallberg Anders, Wolkert Niklas. 1998 *Genres Defined for a Purpose, Fast Clustering, and an Iterative Information Retrieval Interface*, Eighth DELOS Workshop on User Interfaces in Digital Libraries Långholmen, October 1998. pp. 60-66.

Kessler, Brett, Geoff Nunberg, and Heinrich Schütze. 1997. *Automatic detection of text genre*. In 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the conference, 7-12 July, Madrid, p. 32-38.

Kohonen, T. 1995. *Self-organizing Maps*. Berlin; Heidelberg; New York: Springer.

Ljungberg, Jan (1997) *Organizations and Conversation*, Proceedings of IRIS 20 – Department of Informatics, University of Oslo

Miller, Carolyn R. 1994. *Genre as social action*. In: Aviva Freedman and Peter Medway (Eds.) Genre and the new rhetoric. London: Taylor and Francis. 23-42.