Using the IPTC Taxonomy to classify articles automatically for the MD Info taxonomy

Arjan Temmink (294270)

Supervisor: Uzay Kaymak

Bachelor Thesis Economics and Informatics
Major Business Information Systems
FEB33100

August 20, 2010

Abstract

Every day more news articles are becoming available on the Internet. These can come from various sources, like traditional news paper agencies or blogs. Due to this increased amount the cost of classifying them manually have steadily increased during recent years. This research is about classifying articles automatically with the help of certain patterns in the article.

Contents

Chapter 1

Introduction

Nowadays the availability of new information grows exponentially. Every day more and more news articles are added to the Web. These articles can come from various sources, like news agencies or from people who are using the Internet and share their experiences. For example, a person witnesses an accident and decides to write an article about it and publishes it on the Internet. Especially the input of this last group have increased in size in recent years.

So we have an enormous amount of information available, but how can we retrieve the information we are looking for? Obviously, we can't read every article in the database, because this would take days or even months depending on the size of the database. The answer is to make use of a taxonomy. A taxonomy is a hierarchical classification and it can contain several levels. This way we can search for articles selectively. For example if you want to read an article about 'football', you can first look at the general classification 'sport' and then at the sub classification 'football'. For the general idea you can look at figure 1.1. Each article is assigned to a main classification. Then it is further assigned to a sub classification until we have reached the lowest level. So we can search for articles in a quick and selective way by using a taxonomy.

Main Classification
Sub Classification
Other Classification Levels

Figure 1.1: General idea of a taxonomy

MD Info is a company that provides such a taxonomy. It is a company that delivers business information and news to a wide variety of customers. Their taxonomy consists of three levels, the Main Heading, the Sub Heading and the Subject. For economic news it is very specific, while on other areas it is more generalized. With specific is meant that it is very detailed. Every day about 80 articles are coded manually. This is only a small fraction of what is becoming available each day. So now we have a taxonomy that can classify our articles, but it can't keep track with the amount of articles becoming available.

A taxonomy that can do this is the IPTC taxonomy. IPTC is a standard taxonomy from the International Press Telecommunication Council. This agency is a consortium of the world's major news agencies, news publishers and news industry vendors. This taxonomy contains three layers as well, namely the Subject, the SubjectMatter and the SubjectDetail. This taxonomy automatically classifies the articles based on their content. However, it is not very specific concerning economical news. A company that is using this taxonomy is YourNews, which is a part of MD Info.

1.1 Problem statement

As already mentioned, the availability of news grows exponentially. To code all of these articles by hand is a very time-consuming process and it can only be done at a very high cost. If we could process the classification of these articles automatically, a lot more articles can be coded and brought available to the customer. At the moment the taxonomy of MD Info can't classify articles automatically, so we need the help of the IPTC taxonomy. With these two taxonomies we can keep the specificness of the MD Info taxonomy and thanks to the IPTC taxonomy we have a way to classify these articles automatically.

You can contradict this by saying that quality is a lot more important than quantity. This is true, but if you only can process a small fraction of the amount of news available, you might be missing a lot of relevant information.

So in this research we are going to investigate the option of automatically classifying the MD Info taxonomy with the help of the IPTC taxonomy in order to keep track with the availability of news.

1.2 Relevance

In current-day society the Web plays an important role. More and more information can only be found on-line. More and more publishers are deciding to have only an on-line version of their magazines and as a consequence the circulation of news papers is going down. In the current situation MD Info still focuses primarily on off-line content. This is due to the time and the costs it takes to read and classify every article available manually. If MD Info continues on their current way, all this on-line information will be ignored and since the availability of off-line content is decreasing less and less news is becoming available for their customers. More importantly, how can you expect companies to do business with you if you only provide half or even less of the information available.

Another factor is that we are currently living in a rapid-changing environment. News from yesterday can already be outdated. Newspapers are only printed once a day, but with the Internet you don't have this problem. New content can be added every second. In order to keep up, an automatic coding system is needed, which will provide your customers with up-to-date information, so they can make decisions based on all of the news currently available.

With the help of an automatic coding system MD Info can handle more articles each day and as a result providing their customers with more and up-to-date news.

1.3 Research question

The research question is:
'Is it possible to classify articles from the MD Info taxonomy automatically with the help of the IPTC taxonomy?'.
In order to gain an answer to this question, we will consider the following two approaches.

1 *'Is there a direct connection available between certain parts of the taxonomy?'*
This subquestion will be answered by comparing the titles of articles in a certain subject. If a certain article in the IPTC taxonomy has the same title as a article in the MD Info taxonomy, the topics of these articles can be considered 'similar' with respect to a certain threshold.

2 *'Is it possible to provide this similarity by using a word count?'*
If there is no direct connection between the two taxonomies, it may be possible to provide this with the help of a few keywords. If a certain topic in the IPTC taxonomy uses the same words as a topic in the MD Info taxonomy, these topics can be considered 'similar' with respect to a certain threshold.

1.4 Methodology

First of all, we should get familiar with the subject and the tools we are going to use in this research. This is done by searching for relevant articles in academic journals. The relevant articles in this area of research are being discussed in chapter 2.

After we are acquainted with the subject we can start with retrieving all the articles from the MD Info and the YourNews database. Articles are added daily to these databases and by storing them off-line we can ensure that the results of this research will be reproducible. Another advantage is that operations on the data will be a lot quicker if you don't have to retrieve the articles with each operation.

The next step involves providing a classification based on the title of each article. Each title in the MD Info database will be compared to a title in the YourNews database. If they match, we will store in which Subjects these articles were found and we assume that this combination of Subjects will be valid for all the articles in this combination.

In order to provide a classification based on the words that are used in the articles, we have to count them. This will be done with the use of a stemming algorithm. So for each Subject we will get a list of the words in that Subject and how many times these words appear, a so-called 'fingerprint'. This fingerprint is used to define in which subject a certain article belongs.

When this is done, the results will be analyzed and compared with the classification of MD Info. In other words are there any similarities between the taxonomies. Finally, our conclusions are provided.

So this research consists of the following steps:

1. Search and discuss relevant literature.
2. Retrieve the articles from the MD Info and the YourNews database.
3. Provide a classification based on the title.
4. Count the appearance of each word in both databases.
5. Provide a classification based on the fingerprint of each Subject.
6. Analyze and compare the results of the classification algorithm with the original IPTC coding.
7. Provide the conclusions of this research.

1.5 Outline

The remainder of this article is as follows. In chapter 2 the related work in this area is being discussed. Chapter 3 discusses the tools and measures used in this research. In chapter 4 a more detailed overview of the steps taken in this research is provided. Chapter 5 will display the results of this research and chapter 6 will present our conclusions.

Chapter 2

Literature review

In this research we will have to undertake several actions.

1. We also have to retrieve these articles from the web sites of MD Info and YourNews. (Text retrieval)

2. We have to count each word. (Text mining)

3. We have to classify articles according to a certain taxonomy. (Text categorization)

4. We need tools and measurements to analyze the results.

In this chapter a technical background is given concerning these topics.

So first we give more information about the retrieval of the articles. The problem with this is how to retrieve only the information we want. Breuel (2003) proposes a screen-scraping utility. Screen-scraping can be defined as retrieving visible information from a web page. Each web page consists of HTML-code, which can be grouped in a Document Object Model (DOM) tree. Now that the content of the web page is organized in a DOM tree we can easily navigate through this tree and retrieve the relevant data. In figure 2.1 can be seen that the structure of a web page stays the same and only the content changes. Given this information we only have to set up the general structure and as a consequence we can retrieve all of the articles without making changes to the structure. Kosala, Bruynooghe, Van den Bussche & Blockeel (2003) discuss another way to retrieve the relevant information from a web page. Here the data first has to be trained in order to identify the distinguishing context. After the training process the relevant data can be retrieved. Another approach is given by Reis, Golgher, Silva & Laender (2004). They make use of a concept called 'tree edit distance'. This concept is the minimum amount of operations necessarily to transform one tree to another tree. It can be used to identify changing objects in the tree structure which can be the parts where the relevant information is stored. Now that the changing parts are identified, you still have to decide which parts are relevant. They do this by making use of an algorithm that determines the cost of the transformation and if it is relevant. In our research we have followed the approach of Breuel (2003), because this is the most intuitive way and it requires the least amount of operations.

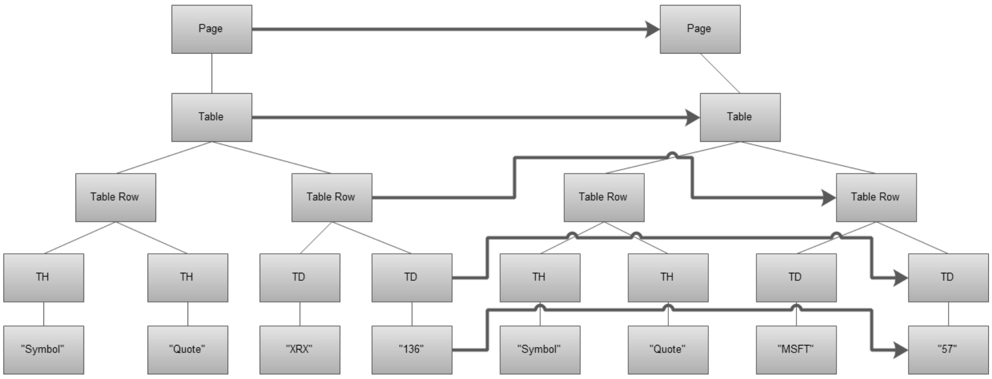

Figure 2.1: Structure HTML in a web page

Second we will discuss text mining. It can be defined as: 'Text mining is a process that derives high-quality information with the help of pattern and trends'. This is essential, because without a proper fingerprint the resulting classification will also be false. In this research text mining will be used to get a relevant word count per topic, a so-called fingerprint. Since text mining is a part of text categorization we will discuss these terms together.

For a proper understanding of this research it is important to know that the taxonomies considered in this research are dynamic. One article can belong to several topics. Sacco (2006) defines a dynamic taxonomy as follows:

> 'A dynamic taxonomy is a taxonomy with a multidimensional classification: a document D can be classified under several topics at any level of abstraction as required.'

To classify the articles according to the IPTC taxonomy a classification algorithm is needed. Sebastiani (2006) wrote an article about automatic text categorization. They distinguish between 'hard' and 'soft' categorization. An example of a hard categorization is the MD Info database. A certain article belongs to a certain topic unconditionally. Soft categorization however assigns scores to a topic like is done in the IPTC taxonomy. As already mentioned the word count is a part of text categorization. By reducing words to their roots, also known as 'stemming', a more meaningful word count can be returned. Han, Karypis & Kumar (1999) propose an algorithm to categorize text. By using a modified version of k-Nearest Neighbors (kNN), Weight Adjusted k-Nearest Neighbors where each word gets a weight reflecting its importance, it can also give good results in documents with a lot of words. Another algorithm based on kNN, the kNN model-based algorithm, is proposed by Guo, Wang, Bell, Bi & Greer (2006). This algorithm combines the kNN classifier with the Rocchio classifier. Where kNN is similarity-based, the Rocchio method is a linear classifier. By combining the strengths of these two classifiers, kNN model-based is taking away some of the drawbacks of these methods. Nefti, Oussalah & Rezgui (2009) uses a completely different method for document categorization, namely fuzzy clustering. This approach looks for similarities between the documents and places a weight on them.

Subramaniam, Nanavati & Mukherjea (2009) wrote an article about how to merge taxonomies. They identify two steps. The first step is to map the taxonomies, so that similar concepts can be identified. The second step is merging or integrating the both taxonomies. This has to be done done with respect to the coherency, the consistency and the redundancy.

Similar research in the area of automatic classification of articles according to the IPTC taxonomy is done by Bacan, Pandzic & Gulija (2005) for the Croatian language. For the classification of the articles, they have used the kNN algorithm in combination with a weighted word count. However, this research will be going one step further by also taking the lower levels of the taxonomy into account and combining the IPTC taxonomy with the taxonomy of MD Info.

The tools and measurements we have used will be discussed in the next chapter.

Chapter 3

Analysis measures

In this chapter we will explain the different tools and measures we are using in this research. We need to have tools that can ensure a good word count and measures that can measure the classification.

If we would just use a word count based on how many times a certain words appears in a Subject, we will get in the top rankings many irrelevant verbs and prepositions. So we will need a tool that gives these words a lower ranking. A tool that can do this is tf-idf (term frequency-inverse document frequency). Tf-idf evaluates how important a certain word is in the Subject.

If we want to know how well a certain classification is, we will need measures to measure its quality. With quality is meant how much of the original dataset is recovered and how well is it classified. There are two widely uses measures in the Information Retrieval field to describe the quality of a search. These are 'recall' and 'precision'. Recall is about how much of the dataset is recovered and precision about how precise the classification is. There is also work written about these measurements tools. Goutte & Gaussier (2005) discusses how well we should trust these values. First they explain these values. Then they show how well these values are by means of a probabilistic framework. Since the same measures will be used, this is a very useful article to understand these concepts.

3.1 Tf-idf

As already mentioned, tf-idf is a tool that evaluates how important a certain word is in the Subject. This importance increases proportionally to the number of times a certain words appears in the Subject, but is offset by the number of articles it appears on. Tf-idf falls into two parts, the term-frequency (tf) and the inverse document frequeny (idf).

The term frequency can be calculated by counting the number of appearance of a certain word i in Subject j divided by the total number of words in that Subject. This can be seen in equation 3.1.

$$Tf_{i,j} = \frac{n}{\Sigma_{i,j} n} \tag{3.1}$$

The inverse document frequency can be calculated by dividing the total number of articles divided by the number of articles a certain word i appears. Then we have to take the logarithm of this result as can be seen in equation 3.2.

$$\texttt{Idf} = log \frac{|D|}{|d : t_i \epsilon D|} \tag{3.2}$$

By multiplying the tf score with the idf score, we obtain the tf-idf score, as can be seen in equation 3.3.

$$\texttt{Tf-idf} = tf * idf \tag{3.3}$$

Word	Word count	Total words	Article Count	Total Articles
overheid	200	5000	40	1000

Table 3.1: Data tf-idf

Let us consider the example in table 3.1. The word ‘overheid’ appears 200 times and our total number of words is 5000. So the term frequency is:

$$Tf_{i,j} = \frac{200}{5000} = 0.04 \tag{3.4}$$

Further is given that this word appears in 40 from the 1000 articles, so the inverse document frequency is:

$$\texttt{Idf} = log \frac{1000}{40} \approx 1.40 \tag{3.5}$$

Now that we know the tf score as the idf score we can calculate the tf-idf score.

$$\texttt{Tf-idf} = 0.04 * 1.40 \approx 0.056 \tag{3.6}$$

How important this word is depends on the scores of the other words in the corpus, but how higher the score the more important the word is.

3.2 General definition of the measurement tools

In this section we will provide a framework on how to use precision and recall. Both measures are described in terms of retrieved documents and relevant documents. The set of Retrieved Relevant Documents can be defined by the intersection of both: $|RetrievedDocuments \cap RelevantDocuments|$. A graphical display can be seen in figure 3.1.

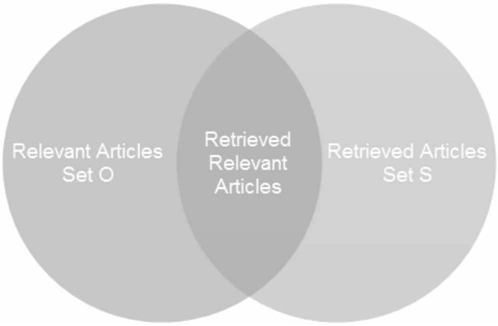

Figure 3.1: general form

Precision is defined by the number of retrieved relevant documents divided by the number retrieved documents.

$$Precision = \frac{|RetrievedDocuments \cap RelevantDocuments|}{|RetrievedDocuments|} \tag{3.7}$$

Recall is defined by the number of retrieved relevant documents divided by the number of relevant documents.

$$Recall = \frac{|RetrievedDocuments \cap RelevantDocuments|}{|RelevantDocuments|} \tag{3.8}$$

For a search to be of good quality we want both precision and recall to be good. A measure that combines the both is the ‘f-score’. The f-score takes the harmonic average of precision and recall. Here is the formula:

$$\texttt{F-score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.9}$$

3.3 Measuring performance for Subjects

In this research for each subject we try to retrieve the articles that belong to that certain subject, by taking the best ten words. These words come from the word count ranked by the tf-idf score. We measure the performance on precision, recall and the f-score. We assume that the set of articles that come with a subject are all the relevant documents. So for a subject X there is a set O with articles, the relevant articles. For each query search there is a corresponding set S with articles, the retrieved articles. For each query, precision and recall shall be calculated as follows:

$$Precision = \frac{|S \cap O|}{|S|} \tag{3.10}$$

$$Recall = \frac{|S \cap O|}{|O|} \tag{3.11}$$

Equation (3.10) and (3.11) are the mathematical counterparts of respectively equation (3.7) and (3.8). With the precision and recall, the F-score can be calculated.

3.4 Overview of the results

When the search for the optimal queries is done, we have a list for each Subject with the precision, recall and f-score for that particular Subject. This list tells us how well articles that belong to that Subject can be retrieved and how precise the classification is by using the IPTC taxonomy.

3.5 Measurement thresholds

Having the highest precision, recall and f-score for all the subjects allows us to set those measures up against thresholds. If a subject passes the threshold, the subject can be marked as ‘similar’.

3.6 Analysis example

Let’s take the sub class ‘Interne bedrijfsaangelegenheden’ as an example. This class consists of 43.777 articles, the set O. 295.341 articles could be retrieved by the IPTC taxonomy, the set S. From these retrieved articles there were 35.625 articles in set O as well. An overview is given by figure 3.2.

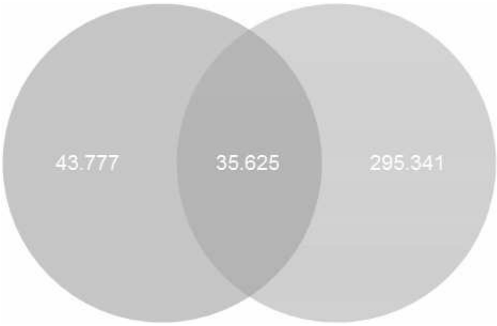

Figure 3.2: sample ‘Interne bedrijfsaangelegenheden’

From these data the precision, recall and f-score can be calculated as follows:

$$Precision = \frac{|S \cap O|}{|S|} = \frac{35625}{295341} = 0.12 \tag{3.12}$$

$$Recall = \frac{|S \cap O|}{|O|} = \frac{35625}{43777} = 0.81 \tag{3.13}$$

$$\texttt{F-score} = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0.12 * 0.81}{0.12 + 0.81} = 0.21 \tag{3.14}$$

The data for this example can be found in listing 3.1.

Listing 3.1: example analysis

```
<Subjects>
  <Name>Interne bedrijfsaangelegenheden</Name>
  <RelevantRetrieved>35625</RelevantRetrieved>
  <Relevant>43777</Relevant>
  <Retrieved>295341</Retrieved>
  <Precision>0.12</Precision>
  <Recall>0.81</Recall>
  <Fscore>0.21</Fscore>
  </Subjects>
```

Chapter 4

Experimental Design

The experimental design consists of five steps that will be done as described in figure 4.1. A detailed explanation of these steps can be found below.

1. Retrieve the articles from the MD Info and the YourNews database.
2. Provide a classification based on the title.
3. Count the appearance of each word in both databases.
4. Provide a classification based on the fingerprint of each Subject.
5. Analyze and compare the results of the classification algorithm with the original IPTC coding.

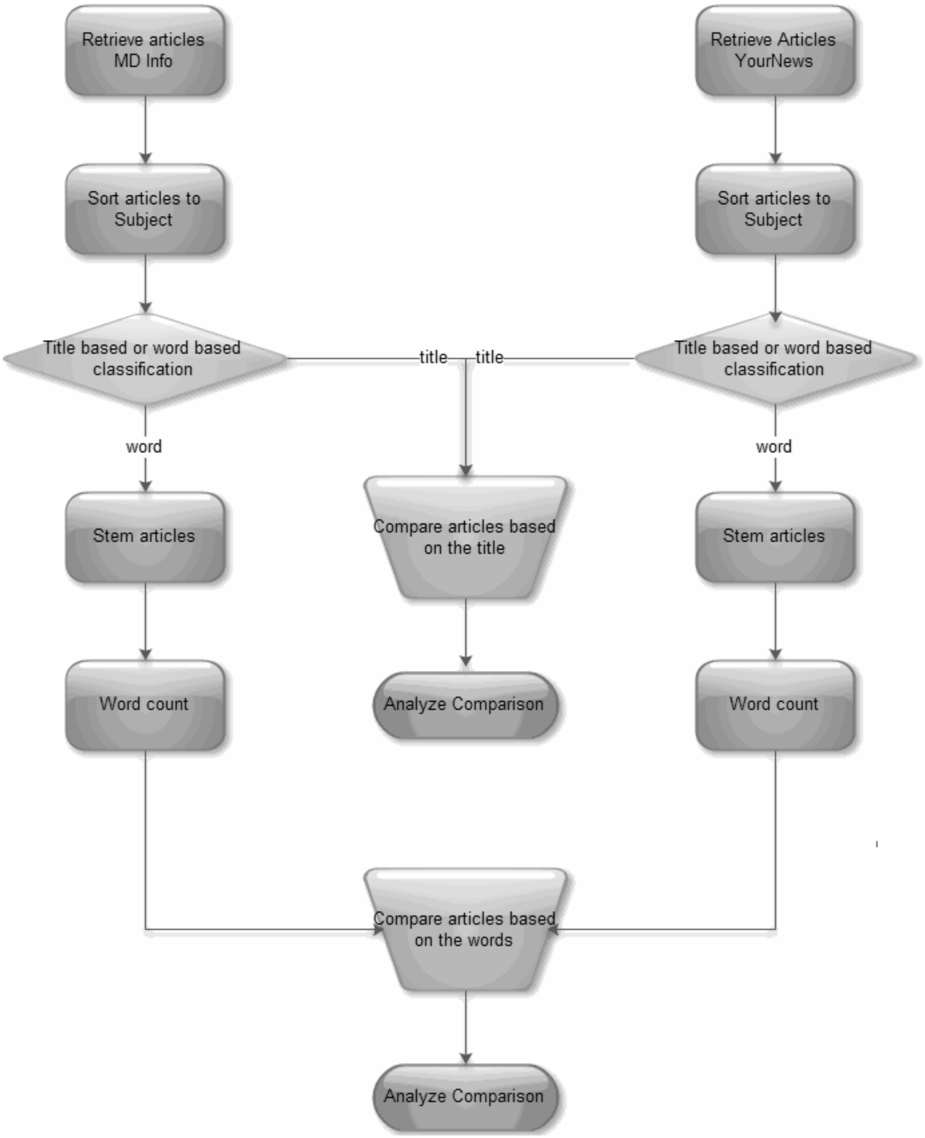
Retrieve articles MD Info
Retrieve Articles YourNews
Sort articles to Subject
Sort articles to Subject
Title based or word based classification
title
title
Title based or word based classification
word
word
Stem articles
Compare articles based on the title
Stem articles
Word count
Analyze Comparison
Word count
Compare articles based on the words
Analyze Comparison

Figure 4.1: Overview of the steps

4.1 Retrieving the articles

The first step in this research is to retrieve the articles from the MD Info and the YourNews databases. We have built a Java application to do this this. Below is a more detailed overview for each database. Since the taxonomies are dynamic, an article can be classified under different topics, even under different sub headings or main headings. In table 4.1, an example can be seen which is based on the article in listing 4.1.

Subject	Main Heading	Sub Heading
Provinciale overheid	Overheid	Overheid
Gemeentelijke overheid (algemeen)	Overheid	Overheid
Rijkswaterstaat	Overheid	Overheidsdiensten
Mobiliteit	Vervoer, verkeer	Verkeer
Decision supportsystemen (ERP, CRM SCM, e.d.)	Automatisering	Software, speciale toepassingen
Samenwerkingen, joint ventures (bedrijfsintern)	Bedrijfsaangelegenheden	Interne bedrijfsaangelegenheden

Table 4.1: Classification articles

4.1.1 MD Info

In the case of MD Info, the application was able to log in to the website of MD Info and retrieving the articles set in a profile[1] with a preset Session ID (SID) number. This SID number was obtained via the website of MD Info manually. By searching for all articles and storing them in a profile, we could obtain the articles with the application. On the 28th of May, the database consisted of 331.800 articles. We have decided to put 10.000 articles in one file, which means we got a total of 34 files. An example of an article in the XML-format can be seen in listing 4.1.

Listing 4.1: sample MD Info: “articles MD Info.xml”

```
<Article>
<ArticleID>935779</ArticleID>
<Title>Alle regionale RWS−verkeerscentrales beschikken over cvms</Title>
<Content>De vijf regionale verkeerscentrales van Rijkswaterstaat beschikken
    sinds 2009 over een Centraal Verkeersregelinstallatie Management Systeem (
    cvms). Naast de regionale diensten van Rijkswaterstaat kunnen ook andere
    regionale en lokale wegbeheerders hier hun voordeel mee doen.
Alle RWS−verkeerscentrales kunnen met de cvms−en de vri‘s (
    VerkeersRegelInstallaties) en tdi‘s (ToeritDoseringsInstallaties) efficient en
    snel beheren. De systemen zijn geschikt voor dynamisch verkeersmanagement
     en ondersteunen de trend dat RWS−verkeerscentrales regelscenario‘s voor
    zowel onderdelen van het autosnelwegennet als delen van het overige
    wegennet inzetten. Zo gaat het cvms van de Regionale
    Verkeersmanagementcentrale Midden Nederland op termijn de vri−systemen
    van grote gemeenten en provincies koppelen. Hiertoe is een regionaal
    samenwerkingsverband van diverse overheden in het leven geroepen.
Een groot voordeel van het ontwikkelen van regionale centrales is dat er zo met
    regiospecifieke kenmerken rekening gehouden kan worden. De complexe
    aansturing van lokale wegen in een regio maakt decentrale verkeerscentrales
    ook min of meer noodzakelijk, naast een ‘centrale‘ centrale. De verschillende
    wegbeheerders binnen een regio kunnen met een gedeeld cvms hun taken
    beter afstemmen. Bij regio−overstijgende zaken of verstoringen met gevolgen
     op landelijke schaal wordt de regie overgenomen door het Verkeerscentrum
    Nederland (VCNL).
</Content>
```

[1] A profile is a saved search query

```
<Subject>Provinciale overheid<SubjectID>27</SubjectID>
</Subject>
<Subject>Gemeentelijke overheid (algemeen)<SubjectID>28</SubjectID>
</Subject>
<Subject>Rijkswaterstaat<SubjectID>2959</SubjectID>
</Subject>
<Subject>Mobiliteit<SubjectID>1852</SubjectID>
</Subject>
<Subject>Decision supportsystemen (ERP, CRM SCM, e.d.)<SubjectID>2059</SubjectID>
</Subject>
<Subject>Samenwerkingen, joint ventures (bedrijfsintern)<SubjectID>2315</SubjectID>
</Subject>
</Article>
```

4.1.2 YourNews

In the case of YourNews, the application also had to take the weight of a subject into account. This weight defines how much an article belongs to a certain subject. In table 4.2, an example can be seen which is based on the article in listing 4.2. These weights are percentages, so the article reflects for 88 percent the category ‘Weernieuws’. Unfortunately, there was no option to retrieve all articles at once, so the articles had to be retrieved per Subject. This also means that some articles are retrieved twice or even more times. So an extra step was required to get only one copy of an article. This step compared the article ID’s of the articles and if the article ID was already in the newly formed database, the second copy was deleted. On the 14th of June, there were 137.797 articles in the database of YourNews. Once again, we have decided to put 10.000 articles into one file, which means we got a total of 14 files. An example of an article in the XML-format can be seen in listing 4.2.

Article	Weernieuws	Overstroming	Weervoorspelling	Natuurramp
Listing 4.2	88	85	84	75

Table 4.2: Subject weight

Listing 4.2: sample YourNews: “articles YourNews.xml”

```
<Article>
<ArticleID>1014499</ArticleID>
<Title>Opnieuw regen in Rio de Janeiro</Title>
<Content>Opnieuw regen in Rio de Janeiro
(Novum/AP) – Enkele uren nadat in het Braziliaanse Rio de Janeiro de hevigste
    regenval ooit was gemeten, begon het woensdag opnieuw te regenen.
```

```
De autoriteiten in Brazilie vrezen voor meer aardverschuivingen en een stijging van het aantal doden.
Dinsdag kwamen al 95 mensen om en raakten meer dan honderd mensen gewond door de aardverschuivingen.
De meeste doden vielen nadat modderstromen de hutten in de sloppenwijken vernielde.
De vorige keer dat er in Rio de Janeiro zoveel regen viel, was op 2 januari 1966.
De Braziliaanse president Luiz Inacio Lula da Silva riep de Brazilianen op om te bidden voor een eind aan de regen.
”Dit␣is␣de␣grootste␣overstroming␣in␣de␣geschiedenis␣van␣Rio␣de␣Janeiro,␣de␣grootste␣hoeveelheid␣regen␣in␣een␣dag.
En␣als␣de␣man␣daarboven␣nerveus␣is␣en␣het␣laat␣regenen,␣dan␣kunnen␣we␣hem␣alleen␣maar␣vragen␣om␣de␣regen␣in␣Rio␣de␣Janeiro␣te␣stoppen,␣zodat␣we␣door␣kunnen␣gaan␣met␣het␣leven␣in␣de␣stad.
</Content>
<Subject>Weernieuws<SubjectID>17003001</SubjectID>
<SubjectWeight>88</SubjectWeight>
</Subject>
<Subject>Overstroming<SubjectID>03005000</SubjectID>
<SubjectWeight>85</SubjectWeight>
</Subject>
<Subject>Weervoorspelling<SubjectID>17001000</SubjectID>
<SubjectWeight>84</SubjectWeight>
</Subject>
<Subject>Natuurramp<SubjectID>03015001</SubjectID>
<SubjectWeight>75</SubjectWeight>
</Subject>
</Article>
```

4.2 Word count

Now that all articles have been retrieved, we can continue with counting the words for each subject in the YourNews database. For this word count we have made use of the stemming algorithm for the Dutch language from Kraaij & Pohlmann (1995). The stemming process can be found in algorithm 1. First is checked if the word is on the Dutch stop word list. This is a list with the most common Dutch words. If it is not on the list, the word is stemmed according to the algorithm.

Algorithm 1 Stemming articles

```
for all articles do
    $Words \leftarrow$ Title and content
    for all words do
        if Word is in stop list then
            Delete word
        else
            Stem word using the algorithm
        end if
    end for
    return The stemmed article
end for
```

Further we have made use of tf-idf to ensure that the more important words have a higher ranking. Otherwise common verbs and prepositions will be ranked high in the list. The word count can be seen in algorithm 2. In line 3 to 13 is counted how many times each word appears in the Subject and in how many articles. Line 14 to 19 calculates for each word how important it is in the Subject. The article percentage is a measure that indicates how important the word is in terms of articles. Suppose a word appears in 4 out of 10 articles. This will make the article percentage 40 %. For the formula you can look at equation 4.1. Finally the words are ranked based on the tf-idf score. For each word we get a number of values. These values can be found in listing 4.3.

$$\texttt{Article percentage} = \frac{|a : t_i \epsilon A|}{|A|} \tag{4.1}$$

Algorithm 2 Count words

```
for all articles do
  $Words \leftarrow$ Title and content
  for all words do
    if Word is in list then
      Increment counter with 1
    else
      Add word to list with count = 1 and article count = 1
    end if
  end for
  if word appeared in this article then
    Increment article count with 1
  end if
end for
for all Words do
  Calculate the article percentage
  Calculate the term frequency
  Calculate the inverse document frequency
  Calculate the tf-idf score
end for
Rank words based on the tf-idf score
return A ranking based on tf-idf per Subject
```

Listing 4.3: sample word

```
<Words>
  <Word>ministeries</Word>
  <Count>9</Count>
  <NumberOfArticles>2</NumberOfArticles>
  <ArticlePercentage>10.53</ArticlePercentage>
  <Tf>0.01</Tf>
  <Idf>0.98</Idf>
  <Tf−idf>0.01</Tf−idf>
  </Words>
```

4.3 Classifying articles

When the words of each subject have been counted, we can classify the articles from MD Info according to the IPTC taxonomy. For every article the stemming algorithm will be applied. An example of the output of this classification can be seen in listing 4.4. The interpretation of the count value depends on which classification is used. For the title based classification the count displays how many times this particular combination is found.

However, for the word based classification the count is an aggregated number of the article percentages.

Listing 4.4: sample output classification

```
<Words>
<Comparison>
  <MDInfo>Werkloosheid</MDInfo>
  <IPTC>Economische indicator</IPTC>
  <Count>17</Count>
  </Comparison>
```

4.3.1 Title based classification

First we considered the idea of comparing the two taxonomies based on the titles of the articles. If the titles are equal, we assume that the articles are the same. If we now look at the classification in both taxonomies, we can assume that these classifications belong together. In algorithm 3 this process can be seen.

Algorithm 3 Title based classification

```
new list of combinations
for all articlesMDInfo do
  for all articlesYourNews do
    if title articlesMDInfo equals title articlesYourNews then
      if Already in list then
        Increment counter of this combination of MD Info and YourNews with 1
      else
        Add combination to list with count = 1
      end if
    end if
  end for
end for
return List of combinations
```

4.3.2 Word based classification

Where the title based classification is dependent on if a title appears in both taxonomies, the word based classification only looks for words in a certain subject. From each subject the best ten words based on the article percentage are taken. This classification process can be seen in algorithm 4. Line 8 to 12 demands some further explanation. Suppose we have the data as listed in table 4.3. Here both words are equal, so we increment the counter with 5.25*4 = 21. If there are more equal words the counter is incremented with their weight. Finally we select the combination with the highest count and add it to the list.

Word MD Info	Article Percentage	Word YourNews	Article Percentage
bank	5.25	bank	4.00

Table 4.3: Data word based classification

Algorithm 4 Word based classification

```
new list of combinations
for all articlesMDInfo do
  Load word count MD Info for this subject
  Select the best 10 based on tf-idf
  for all articlesYourNews do
    Load word count YourNews for this subject
    Select the best 10 based on tf-idf
    for all best 10 MD Info and best 10 YourNews do
      if word articlesMDInfo equals word articlesYourNews then
        Increment counter of this combination of subjects
        with article percentage word MD Info * article
        percentage word YourNews
      end if
    end for
  end for
  Select the combination with the highest count
  Add this combination to the list
end for
return List of combinations
```

4.4 Analysis

Now that we have a list of every combination for these taxonomies, we can analyze this classification. This analysis will be the same for the word based classification as the title based classification. Algorithm 5 contains a detailed description of this process. For each combination the recall, precision and fscore values are calculated.In figure 4.2 can be seen how the retrieved articles are calculated. A certain topic of MD Info can give as a result a topic of YourNews, but this topic can be chosen by multiple topics of MD Info, so the retrieved set is formed by adding the topic size of the several topics.

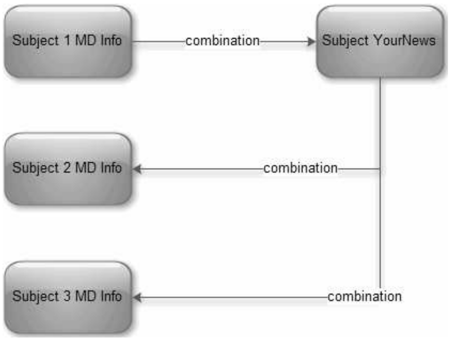

Figure 4.2: example precision

Algorithm 5 Analysis

```
Load list of combinations
for all combinations do
  Load subject MD Info and YourNews
  Retrieve amount of articles in each subject
  Add the amount of articles to the sub heading
  Calculate how many articles are retrieved by the IPTC subject {}See figure
  Calculate recall
  Calculate precision
  Calculate fscore
end for
```

Chapter 5

Results

In this chapter the results of our analysis will be displayed. These results will be sorted on the f-score, since this measure combines the recall and precision measures. Only the best twenty results based on the f-score are shown in this chapter. The complete tables can be found in the appendix. As a reminder, MD Info's taxonomy consists of 60 main headings and 190 sub headings.

5.1 Results for the title based classification

5.1.1 Main heading

The first thing you should notice is that there are only 42 main headings. This means that 18 main headings don't have any connection with the IPTC taxonomy. In other words 30 percent is missing (see equation 5.1).

$$\texttt{Main headings missing} = \frac{18}{60} = 0.30 \tag{5.1}$$

In table 5.1 we can see that 'Horeca' is the best performing main heading with a precision of 1.0 and a recall of 0.16. This means that 16 per cent of the articles in the category 'Horeca' can be retrieved by the IPTC taxonomy and there are no articles present from another category. Another conclusion from this table is that you can't get a high precision without a low recall and vice versa.

Table 5.1: Title based classification - Main Headings

Nr	Name	Precision	Recall	Fscore
1	Horeca	1.0	0.16	0.28
2	Soort artikel	0.16	0.91	0.27
3	Bedrijfsaangelegenheden	0.16	0.85	0.27

Continued on next page

Table 5.1 – continued from previous page

Nr	Name	Precision	Recall	Fscore
4	Algemene economie	0.12	0.65	0.21
5	Onderwijs en onderzoek	0.67	0.09	0.16
6	Post-, tele-, en datacommunicatiediensten	0.09	0.6	0.15
7	Gezondheidszorg	1.0	0.08	0.15
8	Maatschappelijke dienstverlening	0.09	0.2	0.13
9	Banken, bankdiensten	0.07	0.73	0.13
10	Beurswezen, effectenhandel, beleggen	0.06	0.69	0.11
11	Bedrijfseconomie	0.07	0.23	0.1
12	Automatisering	0.06	0.4	0.1
13	Maatschappij	0.04	0.54	0.08
14	Distributie	0.04	0.37	0.08
15	Geneesmiddelen	0.04	0.55	0.07
16	Verzekeringen	0.03	0.39	0.06
17	Overheid	0.04	0.17	0.06
18	Bouwindustrie	0.03	0.71	0.06
19	Petrochemische industrie	0.03	0.61	0.05
20	Personenauto's, tweewielers	0.03	0.6	0.05

5.1.2 Sub heading

The sub headings have the same problem as the main headings, namely that not all sub headings are present. 118 sub headings are missing here. If you calculate this in the same way as (5.1), about 62 percent of the sub headings are missing (see equation 5.2).

$$\texttt{Sub headings missing} = \frac{118}{190} = \frac{59}{85} \approx 0.62 \qquad (5.2)$$

Once again we see that a good precision doesn't lead to a good recall. 'Gezondheidszorg (speciale onderwerpen)' and 'Horeca' have only articles with them that belong to that category, but these articles are only a small fraction of what should have been in those categories. 'Soort artikel' however has retrieved every article that should belong in that category, but also a lot of other articles.

Table 5.2: Title based classification - Sub Headings

Nr	Name	Precision	Recall	Fscore
1	Gezondheidszorg (speciale onderwerpen)	1.0	0.21	0.34
2	Horeca	1.0	0.16	0.28
3	Soort artikel	0.16	1.0	0.27
4	Mediadiensten	0.14	0.53	0.22
5	Interne bedrijfsaangelegenheden	0.12	0.81	0.21
6	Macro-economie	0.11	0.84	0.19

Continued on next page

Table 5.2 – continued from previous page

Nr	Name	Precision	Recall	Fscore
7	Onderwijs en onderzoek	0.67	0.09	0.17
8	Post-, tele- en datacommunicatiediensten	0.09	0.6	0.15
9	Maatschappelijke dienstverlening	0.09	0.2	0.13
10	Banken en Bankdiensten	0.07	0.73	0.13
11	Beurswezen, effectenhandel, beleggen	0.06	0.69	0.11
12	Bedrijfskundige aspecten	0.06	0.34	0.1
13	Maatschappelijke ontwikkelingen	0.05	0.83	0.09
14	Overheid	0.04	0.24	0.07
15	Marketinganalyse en -strategie	0.04	0.37	0.07
16	Geneesmiddelen	0.04	0.62	0.07
17	Distributievormen	0.04	0.62	0.07
18	Verzekeringen	0.03	0.39	0.06
19	Software, speciale toepassingen	0.03	0.41	0.06
20	Luchtvaart	0.03	0.6	0.05

5.2 Results for the word based classification

5.2.1 Main heading

With the word based classification every main heading is retrieved perfectly. Unfortunately it is not very precise. 'Soort artikel' has the best precision rate with a mere 24 per cent.

Table 5.3: Word based classification - Main Heading

Nr	Name	Precision	Recall	Fscore
1	Soort artikel	0.24	1.0	0.39
2	Bedrijfseconomie	0.14	1.0	0.25
3	Bedrijfsaangelegenheden	0.13	1.0	0.23
4	Algemene economie	0.11	1.0	0.2
5	Post-, tele-, en datacommunicatiediensten	0.09	1.0	0.16
6	Vervoer, verkeer	0.07	1.0	0.14
7	Marketing	0.07	1.0	0.14
8	Banken, bankdiensten	0.07	1.0	0.13
9	Overheid	0.06	1.0	0.12
10	Informatieverzorging / informatiediensten / mediadiensten	0.06	1.0	0.12
11	Nederlandse bedrijven	0.06	1.0	0.11
12	Consumentenaangelegenheden	0.06	1.0	0.11
13	Beurswezen, effectenhandel, beleggen	0.06	1.0	0.11
14	Distributie	0.05	1.0	0.1
15	Automatisering	0.05	1.0	0.1
16	Bouwen en wonen	0.04	1.0	0.09
17	Verzekeringen	0.04	1.0	0.08
18	Reclame	0.03	1.0	0.07

Continued on next page

Table 5.3 – continued from previous page

Nr	Name	Precision	Recall	Fscore
19	Maatschappij	0.04	1.0	0.07
20	Gezondheidszorg	0.04	1.0	0.07

5.2.2 Sub heading

Just as with the main headings, every sub heading is also retrieved perfectly. Here we can find a few good scores like 'consumenten (algemeen)' and 'Aankondiging nieuwe producten/diensten' with precision rates over 50 per cent.

Table 5.4: Word based classification - Sub Heading

Nr	Name	Precision	Recall	Fscore
1	consumenten (algemeen)	0.8	1.0	0.89
2	Aankondiging nieuwe producten/diensten	0.55	1.0	0.71
3	Niet van toepassing	0.46	1.0	0.63
4	Soort artikel	0.45	1.0	0.62
5	Bier	0.29	1.0	0.45
6	Milieuaspecten	0.18	1.0	0.31
7	Marketinganalyse en -strategie	0.17	1.0	0.3
8	Consumentengedrag	0.16	1.0	0.28
9	Landbouw en visserij	0.14	1.0	0.24
10	Wet- en regelgeving	0.12	1.0	0.22
11	Interne bedrijfsaangelegenheden	0.12	1.0	0.21
12	Headlines	0.12	1.0	0.21
13	Onderwijs en onderzoek	0.09	1.0	0.17
14	Consumententypologien	0.09	1.0	0.17
15	Post-, tele- en datacommunicatiediensten	0.09	1.0	0.16
16	Bedrijfskundige aspecten	0.08	1.0	0.15
17	Gezondheidszorg (speciale onderwerpen)	0.07	1.0	0.14
18	Koffie, thee	0.08	1.0	0.14
19	Macro-economie	0.08	1.0	0.14
20	Banken en Bankdiensten	0.07	1.0	0.13

5.3 Discussion

The title based classification gives some poor results with many of the categories missing. You could increase this with a bit more flexibility for the title. In this research the titles had to be exactly equal, but you could use some flexibility here, let's say 80 percent has to be equal for example. Precision will likely go down, but recall will go up. Another issue is that the articles from YourNews come primarily from online sources and the articles from MD Info from offline sources. This can explain that many titles are not the same due to a different use of language.

For the word based classification there are two possibilities to improve the precision scores. The first possibility is to look at the amount of words that should be equal in order to assume that two categories are equal. In this research we were satisfied if the outcome was a positive value. You could set a threshold for this value, but to be unbiased you should perform the analysis for each value. After this you could compare the different outcomes. By increasing the value the number of categories we will retrieve will also decline, so the question is how many categories you want to sacrifice in order to have better outcomes. An example of this can be seen in table 5.5. Here 5 words are equal, so with values greater than 5 there will be no similarity between these two topics.

Table 5.5: Top ten words 'Wetenschappelijk onderwijs' and 'Universiteiten en hoge scholen'

Nr	Word	Count	Articles	Word	Count	Articles
1	student	2436	778	opleid	547	180
2	opleid	1112	483	universiteit	1018	367
3	universiteit	2530	910	student	2238	597
4	nederland	1172	572	ov	225	63
5	onderwijs	1010	548	onderwijs	484	263
6	jar	1231	686	ict	200	59
7	hbo	438	204	hebb	537	316
8	hoger	611	358	chipkaart	187	57
9	mer	1009	617	rut	145	27
10	schol	397	183	jar	552	327

The second is the fact that only one third of the IPTC taxonomy is represented in the YourNews database. If there are more options to choose from in order to classify the MD Info articles, it is very likely that precision goes up. So you should find a Dutch database that represents the IPTC taxonomy more.

Chapter 6

Conclusion

In this research we have investigated the possibility of classifying the news articles from the MD Info database automatically. We have considered two approaches: a title based and a word based classification. The rating of these classifications have been calculated with the tools in chapter 3. So for each topic we received three values, the recall, the precision and the f-score.

The title based classification doesn't take all categories into account and the scores of the other categories aren't very high. So we can conclude there is no direct link possible between the MD Info taxonomy and the IPTC taxonomy.

For the word based classification we also had to made a word count. This word count was ranked based on tf-idf, which is described in section 3.1, so we have a fingerprint from each category. As for the results, this classification retrieves all the articles from the MD Info database, but is not very precise. So with the databases used in this research we can say that a word count doesn't provide a good fingerprint. However, as mentioned in section 5.3, with a database that contains more categories the results can be very different.

Bibliography

Bacan, H., Pandzic, I. & Gulija, D. (2005), Automated News Item Categorization, *in* 'Proceedings of the 19th Annual Conference of The Japanese Society for Artificial Intelligence', Citeseer, pp. 251–256.

Breuel, T. (2003), Information extraction from html documents by structural matching, *in* 'Second International Workshop on Web Document Analysis. International Workshop on Web Document Analysis (WDA-2003), located at ICDAR 2003, August 3, Edinburgh'.

Goutte, C. & Gaussier, E. (2005), 'A probabilistic interpretation of precision, recall and F-score, with implication for evaluation', *Advances in Information Retrieval* **3408**, 345–359.

Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. (2006), 'Using kNN model for automatic text categorization', *Soft Computing-A Fusion of Foundations, Methodologies and Applications* **10**(5), 423–430.

Han, E., Karypis, G. & Kumar, V. (1999), 'Text categorization using weight adjusted k-nearest neighbor classification', *Advances in Knowledge Discovery and Data Mining* pp. 53–65.

Kosala, R., Bruynooghe, M., Van den Bussche, J. & Blockeel, H. (2003), Information extraction from web documents based on local unranked tree automaton inference, *in* 'International joint conference on artificial intelligence', Vol. 18, Citeseer, pp. 403–408.

Kraaij, W. & Pohlmann, R. (1995), 'Evaluation of a Dutch stemming algorithm', **1**, 25–43.

Nefti, S., Oussalah, M. & Rezgui, Y. (2009), 'A modified fuzzy clustering for documents retrieval application to document categorization', *Journal of the Operational Research Society* **60**(3), 384–394.

Reis, D., Golgher, P., Silva, A. & Laender, A. (2004), Automatic web news extraction using tree edit distance, *in* 'Proceedings of the 13th international conference on World Wide Web', ACM, p. 511.

Sacco, G. M. (2006), ‘Dynamic taxonomies and guided searches’, *Journal of the American society for information science and technology* **56**(6), 792–796.

Sebastiani, F. (2006), ‘Classification of text, automatic’, *The Encyclopedia of Language and Linguistics* **14**, 457–462.

Subramaniam, L., Nanavati, A. & Mukherjea, S. (2009), ‘Enriching One Taxonomy Using Another’, *IEEE Transactions on Knowledge and Data Engineering* .

Appendix

Table 1: Title based classification - Main Headings

Nr	Name	Precision	Recall	Fscore
1	Horeca	1.0	0.16	0.28
2	Soort artikel	0.16	0.91	0.27
3	Bedrijfsaangelegenheden	0.16	0.85	0.27
4	Algemene economie	0.12	0.65	0.21
5	Onderwijs en onderzoek	0.67	0.09	0.16
6	Post-, tele-, en datacommunicatiediensten	0.09	0.6	0.15
7	Gezondheidszorg	1.0	0.08	0.15
8	Maatschappelijke dienstverlening	0.09	0.2	0.13
9	Banken, bankdiensten	0.07	0.73	0.13
10	Beurswezen, effectenhandel, beleggen	0.06	0.69	0.11
11	Bedrijfseconomie	0.07	0.23	0.1
12	Automatisering	0.06	0.4	0.1
13	Maatschappij	0.04	0.54	0.08
14	Distributie	0.04	0.37	0.08
15	Geneesmiddelen	0.04	0.55	0.07
16	Verzekeringen	0.03	0.39	0.06
17	Overheid	0.04	0.17	0.06
18	Bouwindustrie	0.03	0.71	0.06
19	Petrochemische industrie	0.03	0.61	0.05
20	Personenauto's, tweewielers	0.03	0.6	0.05
21	Marketing	0.04	0.07	0.05
22	Informatieverzorging / informatiediensten / mediadiensten	0.03	0.28	0.05
23	Elektrotechnische industrie	0.03	0.66	0.05
24	Consumentenaangelegenheden	0.03	0.25	0.05
25	Bouwen en wonen	0.03	0.34	0.05
26	Vervoer, verkeer	0.03	0.16	0.04
27	Public relations	0.02	0.61	0.03
28	Nederlandse bedrijven	0.02	0.1	0.03
29	Metaalindustrie	0.01	0.39	0.03
30	Media	0.02	0.23	0.03
31	Consumentenelektronica	0.02	0.38	0.03
32	Textiel	0.01	0.2	0.02
33	Grafische industrie	0.01	0.43	0.02
34	Delfstoffen, grondstoffen, energiebronnen	0.01	0.28	0.02
35	Buitenlandse bedrijven	0.01	0.09	0.02

Continued on next page

Table 1 – continued from previous page

Nr	Name	Precision	Recall	Fscore
36	Woninginrichting	0.0	0.16	0.01
37	Transportmiddelenindustrie	0.01	0.29	0.01
38	Levensmiddelen	0.01	0.09	0.01
39	Commerciele dienstverlening n.e.g.	0.01	0.07	0.01
40	Agrarische sector	0.0	0.08	0.01
41	Vakantie, recreatie, sport, kunst en amusement	0.0	0.02	0.0
42	Kantoor- en bedrijfsbenodigdheden	0.0	0.13	0.0

Table 2: Title based classification - Sub Headings

Nr	Name	Precision	Recall	Fscore
1	Gezondheidszorg (speciale onderwerpen)	1.0	0.21	0.34
2	Horeca	1.0	0.16	0.28
3	Soort artikel	0.16	1.0	0.27
4	Mediadiensten	0.14	0.53	0.22
5	Interne bedrijfsaangelegenheden	0.12	0.81	0.21
6	Macro-economie	0.11	0.84	0.19
7	Onderwijs en onderzoek	0.67	0.09	0.17
8	Post-, tele- en datacommunicatiediensten	0.09	0.6	0.15
9	Maatschappelijke dienstverlening	0.09	0.2	0.13
10	Banken en Bankdiensten	0.07	0.73	0.13
11	Beurswezen, effectenhandel, beleggen	0.06	0.69	0.11
12	Bedrijfskundige aspecten	0.06	0.34	0.1
13	Maatschappelijke ontwikkelingen	0.05	0.83	0.09
14	Overheid	0.04	0.24	0.07
15	Marketinganalyse en -strategie	0.04	0.37	0.07
16	Geneesmiddelen	0.04	0.62	0.07
17	Distributievormen	0.04	0.62	0.07
18	Verzekeringen	0.03	0.39	0.06
19	Software, speciale toepassingen	0.03	0.41	0.06
20	Luchtvaart	0.03	0.6	0.05
21	Petrochemische industrie	0.03	0.61	0.05
22	Personenauto's	0.03	0.68	0.05
23	Elektrotechnische industrie	0.03	0.66	0.05
24	Energiebronnen	0.03	0.2	0.05
25	Bouwen en wonen	0.03	0.34	0.05
26	Headlines	0.02	1.0	0.05
27	Informatiediensten	0.02	0.35	0.04
28	Bouw	0.02	0.85	0.04
29	Bedrijfseconomische aspecten	0.02	0.36	0.04
30	Niet van toepassing	0.02	1.0	0.04
31	Public relations (PR)	0.02	0.61	0.03
32	Nederlandse bedrijven	0.02	0.1	0.03
33	Audiovisuele media	0.02	0.39	0.03
34	Milieuaspecten	0.01	1.0	0.03

Continued on next page

Table 2 – continued from previous page

Nr	Name	Precision	Recall	Fscore
35	Consumentenelektronica	0.01	0.69	0.03
36	Tijdbesteding	0.02	0.47	0.03
37	Consumentengedrag	0.02	0.71	0.03
38	Hardware, randapparatuur, datacommunicatieapp., opslagmedia	0.01	0.46	0.03
39	Inkomens en lonen	0.02	0.32	0.03
40	Kleding	0.01	0.26	0.02
41	Actueel nieuws	0.01	0.24	0.02
42	Basis metaalindustrie	0.01	1.0	0.02
43	Indeling marketing	0.01	0.05	0.02
44	Human interest	0.01	0.61	0.02
45	Grafische industrie	0.01	0.43	0.02
46	Aspecten van de detailhandel	0.01	0.21	0.02
47	Buitenlandse bedrijven	0.01	0.09	0.02
48	Bouwindustrie	0.01	0.51	0.02
49	Aspecten van automatisering	0.01	0.28	0.02
50	Buitenlandse handel	0.01	0.22	0.02
51	Woninginrichting	0.0	0.18	0.01
52	Vliegtuigbouw en ruimtevaart	0.01	0.73	0.01
53	Politiek	0.01	0.1	0.01
54	Machine-industrie	0.01	0.32	0.01
55	Milieuvraagstukken	0.01	0.08	0.01
56	Criminaliteit	0.01	0.48	0.01
57	Voedingsmiddelen n.e.g.	0.0	0.18	0.01
58	Speciale voeding	0.0	0.42	0.01
59	Aspecten informatieverzorging	0.0	0.41	0.01
60	Delfstoffenwinning/-exploratie, Grondstoffen	0.01	0.38	0.01
61	Koopproces, koopgedrag	0.01	0.2	0.01
62	Economische aspecten consumentengedrag	0.01	0.3	0.01
63	Consumententypologien	0.01	0.03	0.01
64	Consumentenomgeving	0.0	0.4	0.01
65	Commercile dienstverlening	0.01	0.07	0.01
66	Vakanties	0.0	0.09	0.0
67	Kantoortechniek (excl. kantoorautomatisering)	0.0	0.64	0.0
68	Printed	0.0	0.02	0.0
69	Computers, randapparatuur en software	0.0	0.13	0.0
70	Arbeidsaspecten	0.0	0.0	0.0
71	Biologische landbouw	0.0	0.98	0.0
72	Akkerbouw	0.0	0.36	0.0

Table 3: Word based classification - Main Heading

Nr	Name	Precision	Recall	Fscore
1	Soort artikel	0.24	1.0	0.39
2	Bedrijfseconomie	0.14	1.0	0.25
3	Bedrijfsaangelegenheden	0.13	1.0	0.23

Continued on next page

Table 3 – continued from previous page

Nr	Name	Precision	Recall	Fscore
4	Algemene economie	0.11	1.0	0.2
5	Post-, tele-, en datacommunicatiediensten	0.09	1.0	0.16
6	Vervoer, verkeer	0.07	1.0	0.14
7	Marketing	0.07	1.0	0.14
8	Banken, bankdiensten	0.07	1.0	0.13
9	Overheid	0.06	1.0	0.12
10	Informatieverzorging / informatiediensten / mediadiensten	0.06	1.0	0.12
11	Nederlandse bedrijven	0.06	1.0	0.11
12	Consumentenaangelegenheden	0.06	1.0	0.11
13	Beurswezen, effectenhandel, beleggen	0.06	1.0	0.11
14	Distributie	0.05	1.0	0.1
15	Automatisering	0.05	1.0	0.1
16	Bouwen en wonen	0.04	1.0	0.09
17	Verzekeringen	0.04	1.0	0.08
18	Reclame	0.03	1.0	0.07
19	Maatschappij	0.04	1.0	0.07
20	Gezondheidszorg	0.04	1.0	0.07
21	Vakantie, recreatie, sport, kunst en amusement	0.03	1.0	0.06
22	Personenauto's, tweewielers	0.03	1.0	0.06
23	Buitenlandse bedrijven	0.03	1.0	0.06
24	Onderwijs en onderzoek	0.03	1.0	0.05
25	Media	0.03	1.0	0.05
26	Delfstoffen, grondstoffen, energiebronnen	0.03	1.0	0.05
27	Consumentenelektronica	0.02	1.0	0.05
28	Commerciele dienstverlening n.e.g.	0.03	1.0	0.05
29	Bouwindustrie	0.02	1.0	0.05
30	Agrarische sector	0.03	1.0	0.05
31	Textiel	0.02	1.0	0.04
32	Nutsbedrijven	0.02	1.0	0.04
33	Levensmiddelen	0.02	1.0	0.04
34	Geneesmiddelen	0.02	1.0	0.04
35	Elektrotechnische industrie	0.02	1.0	0.04
36	Dranken	0.02	1.0	0.04
37	Agrarische sector consumentenmarkt	0.02	1.0	0.04
38	Transportmiddelenindustrie	0.01	1.0	0.03
39	Petrochemische industrie	0.02	1.0	0.03
40	Metaalindustrie	0.02	1.0	0.03
41	Horeca	0.02	1.0	0.03
42	Chemische industrie	0.02	1.0	0.03
43	Woninginrichting	0.01	1.0	0.02
44	Public relations	0.01	1.0	0.02
45	Maatschappelijke dienstverlening	0.01	1.0	0.02
46	Durables	0.01	1.0	0.02
47	Cosmetica en drogisterijartikelen	0.01	1.0	0.02
48	Zuivel, zuivelproducten	0.0	1.0	0.01
49	Verpakkingsindustrie	0.0	1.0	0.01
50	Textielindustrie	0.0	1.0	0.01
51	Sponsoring	0.01	1.0	0.01
52	Rookwaren, rookartikelen	0.0	1.0	0.01

Continued on next page

Table 3 – continued from previous page

Nr	Name	Precision	Recall	Fscore
53	Marktonderzoek	0.0	1.0	0.01
54	Kantoor- en bedrijfsbenodigdheden	0.01	1.0	0.01
55	Grafische industrie	0.01	1.0	0.01
56	Doe-het-zelf, foto en film	0.01	1.0	0.01
57	Papier-, karton- en golfkartonindustrie	0.0	1.0	0.0
58	Leerindustrie	0.0	1.0	0.0
59	Instrumenten- en optische industrie	0.0	1.0	0.0
60	Houtbewerkende en houtverwerkende industrie	0.0	1.0	0.0

Table 4: Word based classification - Sub Heading

Nr	Name	Precision	Recall	Fscore
1	consumenten (algemeen)	0.8	1.0	0.89
2	Aankondiging nieuwe producten/diensten	0.55	1.0	0.71
3	Niet van toepassing	0.46	1.0	0.63
4	Soort artikel	0.45	1.0	0.62
5	Bier	0.29	1.0	0.45
6	Milieuaspecten	0.18	1.0	0.31
7	Marketinganalyse en -strategie	0.17	1.0	0.3
8	Consumentengedrag	0.16	1.0	0.28
9	Landbouw en visserij	0.14	1.0	0.24
10	Wet- en regelgeving	0.12	1.0	0.22
11	Interne bedrijfsaangelegenheden	0.12	1.0	0.21
12	Headlines	0.12	1.0	0.21
13	Onderwijs en onderzoek	0.09	1.0	0.17
14	Consumententypologien	0.09	1.0	0.17
15	Post-, tele- en datacommunicatiediensten	0.09	1.0	0.16
16	Bedrijfskundige aspecten	0.08	1.0	0.15
17	Gezondheidszorg (speciale onderwerpen)	0.07	1.0	0.14
18	Koffie, thee	0.08	1.0	0.14
19	Macro-economie	0.08	1.0	0.14
20	Banken en Bankdiensten	0.07	1.0	0.13
21	Maatschappelijke ontwikkelingen	0.06	1.0	0.12
22	Nederlandse bedrijven	0.06	1.0	0.11
23	Marketing mix	0.06	1.0	0.11
24	Frisdranken	0.06	1.0	0.11
25	Beurswezen, effectenhandel, beleggen	0.06	1.0	0.11
26	Arbeidsaspecten	0.06	1.0	0.11
27	Actueel nieuws	0.05	1.0	0.1
28	Overheid	0.05	1.0	0.1
29	Bouwen en wonen	0.04	1.0	0.09
30	Verzekeringen	0.04	1.0	0.08
31	Politieke partijen	0.04	1.0	0.08
32	Aspecten informatieverzorging	0.04	1.0	0.07
33	Distributievormen	0.03	1.0	0.07

Continued on next page

Table 4 – continued from previous page

Nr	Name	Precision	Recall	Fscore
34	Tijdbesteding	0.04	1.0	0.07
35	Software, speciale toepassingen	0.04	1.0	0.07
36	Luchtvaart	0.03	1.0	0.06
37	Reclamevormen	0.03	1.0	0.06
38	Informatiediensten	0.03	1.0	0.06
39	Aspecten van de detailhandel	0.03	1.0	0.06
40	Buitenlandse bedrijven	0.03	1.0	0.06
41	Buitenlandse handel	0.03	1.0	0.06
42	Vervoer	0.03	1.0	0.05
43	Personenauto's	0.03	1.0	0.05
44	Printed	0.03	1.0	0.05
45	Gezondheidszorg	0.03	1.0	0.05
46	Gebruiksaspecten	0.03	1.0	0.05
47	Economische aspecten consumentengedrag	0.03	1.0	0.05
48	Consumentenomgeving	0.03	1.0	0.05
49	Commercile dienstverlening	0.03	1.0	0.05
50	Bedrijfseconomische aspecten	0.03	1.0	0.05
51	Inkomens en lonen	0.02	1.0	0.05
52	Hulpbedrijven (verkeer)	0.02	1.0	0.04
53	Promotions	0.02	1.0	0.04
54	Levensmiddelen	0.02	1.0	0.04
55	Elektrotechnische industrie	0.02	1.0	0.04
56	Ontwikkelingen in distributie en detailhandel	0.02	1.0	0.04
57	Hardware, randapparatuur, datacommunicatieapp., opslagmedia	0.02	1.0	0.04
58	Verkeer	0.02	1.0	0.03
59	Scheepvaart	0.02	1.0	0.03
60	Railvervoer	0.01	1.0	0.03
61	Hulpbedrijven (vervoer)	0.02	1.0	0.03
62	Kleding	0.02	1.0	0.03
63	Reclamemiddelen	0.02	1.0	0.03
64	Petrochemische industrie	0.02	1.0	0.03
65	Politiek	0.02	1.0	0.03
66	Audiovisuele media	0.02	1.0	0.03
67	Indeling marktonderzoek	0.01	1.0	0.03
68	Merken	0.01	1.0	0.03
69	Indeling marketing	0.02	1.0	0.03
70	Onrust, conflicten en oorlogen	0.01	1.0	0.03
71	Mediadiensten	0.02	1.0	0.03
72	Horeca	0.02	1.0	0.03
73	Geneesmiddelen	0.02	1.0	0.03
74	Winkelinrichting	0.02	1.0	0.03
75	Energiebronnen	0.02	1.0	0.03
76	Delfstoffenwinning/-exploratie, Grondstoffen	0.01	1.0	0.03
77	Babyverzorging	0.02	1.0	0.03
78	Consumentenelektronica	0.01	1.0	0.03
79	Koopproces, koopgedrag	0.01	1.0	0.03
80	Bouwindustrie	0.01	1.0	0.03
81	Bouw	0.02	1.0	0.03
82	Bedrijfstypen	0.02	1.0	0.03

Continued on next page

Table 4 – continued from previous page

Nr	Name	Precision	Recall	Fscore
83	Aspecten van automatisering	0.01	1.0	0.03
84	Vis en visproducten	0.02	1.0	0.03
85	Woninginrichting	0.01	1.0	0.02
86	Wegvervoer	0.01	1.0	0.02
87	Vakanties	0.01	1.0	0.02
88	Sport en spel	0.01	1.0	0.02
89	Kunst, amusement	0.01	1.0	0.02
90	Reclame-onderzoek	0.01	1.0	0.02
91	Reclamebureaus	0.01	1.0	0.02
92	Public relations (PR)	0.01	1.0	0.02
93	Elektriciteitsbedrijven	0.01	1.0	0.02
94	Machine-industrie	0.01	1.0	0.02
95	Printed	0.01	1.0	0.02
96	Milieuvraagstukken	0.01	1.0	0.02
97	Demografie	0.01	1.0	0.02
98	Maatschappelijke dienstverlening	0.01	1.0	0.02
99	Zoetwaren	0.01	1.0	0.02
100	Zoet broodbeleg	0.01	1.0	0.02
101	Voedingsmiddelen n.e.g.	0.01	1.0	0.02
102	Huid- en gelaatsverzorging	0.01	1.0	0.02
103	Computers, randapparatuur en software	0.01	1.0	0.02
104	Chemische industrie	0.01	1.0	0.02
105	Aardappelen, groente en fruit	0.01	1.0	0.02
106	Veeteelt	0.01	1.0	0.02
107	Tuinbouw	0.01	1.0	0.02
108	Biologische landbouw	0.01	1.0	0.02
109	Zuivelproducten	0.0	1.0	0.01
110	Verpakkingsindustrie	0.0	1.0	0.01
111	Recreatie	0.01	1.0	0.01
112	Vliegtuigbouw en ruimtevaart	0.01	1.0	0.01
113	Scheepsbouw- en scheepsreparatiebedrijven	0.0	1.0	0.01
114	Bedrijfsvervoer	0.01	1.0	0.01
115	Textielindustrie	0.0	1.0	0.01
116	Schoenen en lederwaren	0.01	1.0	0.01
117	Sponsoring	0.01	1.0	0.01
118	Rookwaren, rookartikelen	0.0	1.0	0.01
119	Regulering van de reclame	0.0	1.0	0.01
120	Indeling reclame	0.0	1.0	0.01
121	Adverteerders	0.01	1.0	0.01
122	Tweewielers	0.0	1.0	0.01
123	Nutsbedrijven	0.01	1.0	0.01
124	Gasdistributiebedrijven	0.0	1.0	0.01
125	Metaalproductenindustrie (excl. machines, transportmiddelen)	0.0	1.0	0.01
126	Defensie-industrie	0.0	1.0	0.01
127	Basis metaalindustrie	0.01	1.0	0.01
128	Onderzoeksmethoden	0.01	1.0	0.01
129	Onderzoek nieuwe producten/diensten	0.01	1.0	0.01
130	Criminaliteit	0.0	1.0	0.01
131	Speciale voeding	0.01	1.0	0.01

Continued on next page

Table 4 – continued from previous page

Nr	Name	Precision	Recall	Fscore
132	Snackproducten	0.0	1.0	0.01
133	Onderleggers	0.0	1.0	0.01
134	Kantoorverbruiksartikelen	0.0	1.0	0.01
135	Kantoor- en bedrijfsinrichting	0.0	1.0	0.01
136	Grafische industrie	0.01	1.0	0.01
137	Speelgoed, spellen, hobby-artikelen	0.0	1.0	0.01
138	Gemengde branche	0.0	1.0	0.01
139	Zwak alcoholische dranken	0.0	1.0	0.01
140	Sterk alcoholische dranken	0.0	1.0	0.01
141	Doe het zelf (incl. metaalwaren en gereedschappen)	0.01	1.0	0.01
142	Detailhandelsorganisatievormen	0.0	1.0	0.01
143	Haarverzorging	0.0	1.0	0.01
144	Drogisterijartikelen n.e.g.	0.0	1.0	0.01
145	Cosmetica en drogisterij artikelen	0.0	1.0	0.01
146	Badproducten, deodorants, toiletzeep	0.0	1.0	0.01
147	Elektrische huishoudelijke apparatuur	0.01	1.0	0.01
148	Kunststof- en rubberverwerkende industrie	0.0	1.0	0.01
149	Biotechnologie	0.01	1.0	0.01
150	Vlees en vleeswaren (incl. wild en gevogelte)	0.01	1.0	0.01
151	Petfoods en dierenbenodigdheden	0.0	1.0	0.01
152	Bloemen, planten en tuinartikelen	0.01	1.0	0.01
153	Visserij, viskwekerij, visproducten	0.0	1.0	0.01
154	Bosbouw	0.0	1.0	0.01
155	Akkerbouw	0.0	1.0	0.01
156	Verwarming, klimaatbeheersing	0.0	1.0	0.0
157	Pijpleidingvervoer	0.0	1.0	0.0
158	Huishoudtextiel	0.0	1.0	0.0
159	Fournituren	0.0	1.0	0.0
160	Papier-, karton- en golfkartonindustrie	0.0	1.0	0.0
161	Overheidsdiensten	0.0	1.0	0.0
162	Wetenschap en techniek	0.0	1.0	0.0
163	Watervoorziening	0.0	1.0	0.0
164	Warmtevoorziening	0.0	1.02	0.0
165	Marktonderzoekstypen	0.0	1.0	0.0
166	Marktonderzoekbureaus	0.0	1.0	0.0
167	Marketing(advies)bureaus	0.0	1.0	0.0
168	Weer	0.0	1.0	0.0
169	Sport	0.0	1.0	0.0
170	Ongelukken en rampen	0.0	1.0	0.0
171	Human interest	0.0	1.0	0.0
172	Diepvriesproducten	0.0	1.0	0.0
173	Leerindustrie	0.0	1.0	0.0
174	Kantoortechniek (excl. kantoorautomatisering)	0.0	1.0	0.0
175	Instrumenten- en optische industrie	0.0	1.0	0.0
176	Houtbewerkende en houtverwerkende industrie	0.0	1.0	0.0
177	Indeling geneesmiddelen	0.0	1.0	0.0
178	Hulpmiddelen	0.0	1.0	0.0
179	Uurwerken, sieraden en edelmetaal	0.0	1.0	0.0
180	Optische artikelen	0.0	1.0	0.0

Continued on next page

Table 4 – continued from previous page

Nr	Name	Precision	Recall	Fscore
181	Muziekinstrumenten	0.0	1.0	0.0
182	Durables n.e.g.	0.0	1.0	0.0
183	Dranken	0.0	1.0	0.0
184	Foto, film	0.0	1.0	0.0
185	Winningswijze	0.0	1.0	0.0
186	Parfums, reuk- en toiletwaters	0.0	1.0	0.0
187	Mondverzorging	0.0	1.01	0.0
188	Herencosmetica	0.0	1.0	0.0
189	Disposables	0.0	1.0	0.0
190	Decoratieve cosmetica	0.0	1.0	0.0