

Report on strategies for collecting URLs

Deliverable Number: 6.6.1.1.2-4



Version: 1.0

Date: 2008-05-14

Author: Morten Goodwin Olsen, Nils Ulltveit-Moe and Mikael Snaprud

Dissemination Level: Public

Status: FINAL

License:

This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/2.5/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

This document consists of 25 pages including this cover (2 pages).

Abstract

This document outlines strategies for categorising NACE and NUTS categories. We present both manual and automatic approaches, including suggestions based on machine learning.

Our experiments indicate that the most reliable approach for NACE categorisation is a classification approach based on term frequencies. A small proof of concept implementation provide results with an accuracy of 100% for NACE categorisation. A disadvantage with the current implementation is that the algorithm needs to be trained for each language.

In contrast, for NUTS categorisation, our tests indicate manual classification is still most efficient.

Version Control

<i>Version</i>	<i>Status</i>	<i>Date</i>	<i>Change</i>	<i>Author</i>
0.1	DRAFT	2008-05-14	First draft – Machine Learning based approach. General Cleanup	Morten Goodwin Olsen
0.2	DRAFT	2008-05-14	General cleanup	Morten Goodwin Olsen
0.3	DRAFT	2008-05-14	General cleanup – integrated comments from Mikael Snaprud	Morten Goodwin Olsen
0.4	DRAFT	2008-05-14	Integrated comments from Nils Ulltveit-Moe	Morten Goodwin Olsen
0.5	DRAFT	2008-05-14	Small adjustments, formatting.	Mikael Snaprud
0.6	DRAFT	2008-05-14	Integrated typography issues from Nils Ulltveit-Moe	Morten Goodwin Olsen
1.0	FINAL	2008-05-14	Made final	Morten Goodwin Olsen

Table of Contents

1 Introduction.....	6
1.1 Brief project description.....	6
1.2 Scope of this document	6
1.3 Objective.....	6
1.4 Related work.....	6
1.5 Overview and readers instructions.....	8
2 Sources of NACE and NUTS codes.....	9
2.1 The European Business Register	9
2.1.1 National business registers.....	9
3 Automatic NACE / NUTS categorisation.....	11
3.1 NACE and NUTS classifications based on fields in the Cag Gemini URL set.....	11
3.1.1 NACE categorisation.....	12
3.1.2 NUTS categorisation.....	12
3.2 Geographical lookup.....	13
3.2.1 Whois.....	13
3.2.2 Hostip.info	13
3.2.3 NUTS Location from location hints in web site.....	13
4 Automatic mMachine learning based approach for NACE/NUTS categorisation.....	14
4.1 NACE categorisation using Nearest Neighbour based on the term frequencies.....	14
4.1.1 Term frequencies for NACE documents.....	15
4.1.2 Training - NACE.....	16
4.1.3 Classification - NACE.....	17
4.1.4 Proof of concept implementation - NACE.....	17
4.2 NUTS classification using a classification tree.....	19
4.2.1 Term frequencies - NUTS.....	21
4.2.2 Decision tree - NUTS.....	22
4.2.3 Classification tree - NUTS.....	22
4.2.4 Proof of concept implementation - NUTS.....	23
5 Manual Categorisation of NACE and NUTS.....	24
5.1 Manual procedure for categorising a web site into NACE	24
5.2 Manual procedure for categorising a web site into NUTS	25
6 Comparing approaches.....	25
6.1 NACE categorisation.....	25
6.2 NUTS categorisation.....	26
7 Conclusion.....	28
8 References.....	28

1 Introduction

1.1 Brief project description

The main objective of the EIAO project, is to create a prototype Observatory for large-scale accessibility evaluations. The URL repository is an important part of EIAO, as it maintains the main inventory of web pages and web sites to be used and updated by the crawler and sampler.

1.2 Scope of this document

This document covers strategies to categorise web sites reliably into geographic regions (NUTS - [1]) and business sectors (NACE - [2]).

The list of URLs used by Capgemini for the 20 services ([3]) reports is used as an example basis for this approach.

Please note that previous versions of this deliverable deal with where to obtain existing lists of URLs. To allow for some comparisons with previous evaluation results and to have a list of URLs selected according to the same criteria in all countries we have chosen to use the list of URLs prepared by Capgemini The User Challenge Benchmarking The Supply Of Online Public Services, 7th Measurement | September 2007 .¹

1.3 Objective

The objective of this document, is to present ideas and approaches for categorising web sites into NUTS and NACE categories.

1.4 Related work

The URL directory is related to:

- D6.6.1.1.2-3 : **Report on strategies for collecting URLs**: The objective is to describe how the EIAO URL repository can be extended with sufficient URLs so that we can meet the requirement of crawling 10 000 different public European web sites each month, and show results from accessibility assessments of these sites aggregated over both branch categories (a subset of NACE) and geographical regions (NUTS).
- The **REGULATION (EC) No 808/2004 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 April 2004 concerning Community statistics on the information society** aims to establish a common framework for the systematic production of Community statistics on the information society. Module 1 covers enterprises and the information society, and mentions breakdowns by NACE heading and by region.

¹ http://ec.europa.eu/information_society/eeurope/i2010/docs/benchmarking/egov_benchmark_2007.pdf

1.5 Overview and readers instructions

Figure 1.1 presents an overview of the URL handling including *Input*, *Use* and *Maintenance* of the URL repository. In this document, we focus on the challenges of categorisation.

In chapter 1 we present a general introduction. Chapter 2 includes available sources for NACE and NUTS categories. This is followed in chapter 3 by Source Based Automatic NACE / NUTS categorisation and in chapter 4 a Machine Learning based approach for NACE and NUTS categorisation. Furthermore, we present strategies for manual categorisation of web sites in chapter 5. Finally, we compare the different approaches in chapter 6 and present a conclusion in chapter 7.

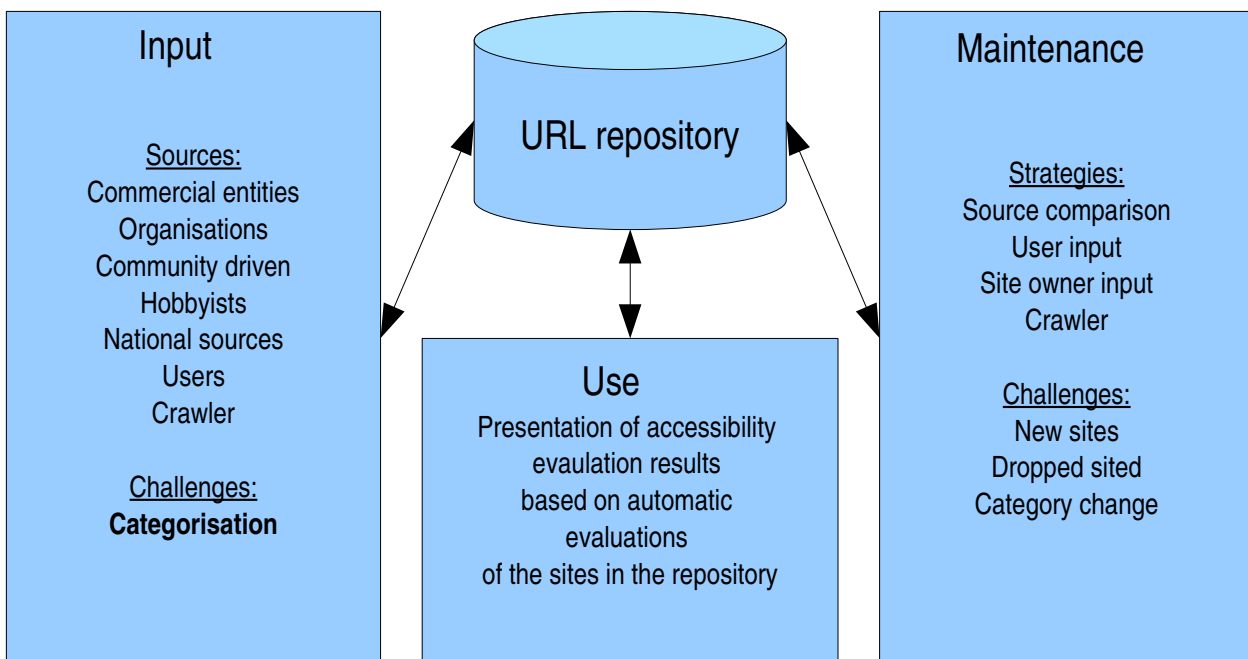


Figure 1.1: Issues related to the URL repository

2 Sources of NACE and NUTS codes

EIAO uses two categories to classify websites : NUTS ([1]) and NACE ([2]).

The NUTS categorisation scheme corresponds to the geographical location of the organisation behind a website, and the NACE categorisation scheme describes the business sector of the organisation behind the website.

There are several existing registries that store information about organisations and their statistical categories. They exist both on European and on national level. Different approaches to extract category information are discussed in the following sections.

In this chapter we present available sources for NUTS and NACE categories.

2.1 The European Business Register

NACE categorisation of businesses are available from the European Business Register (EBR - [4]). EBR is a network of business registers, kept by the registration authorities. 19 of 27 member states are partners in the EBR network, so the database does not yet have complete European coverage. Information in the EBR is copyrighted to the national authorities under the database directive. Access to business register data is sold on a per-company basis with a restrictive license that only allows personal use, and not redistribution.

Getting permission to use these data would require approval from each individual European country and/or businesses that own the underlying data. From an economical perspective, the price of searches for basic business information in the EBR ranges from 2,8 – 7.5 Euro. As an example, categorising 10 000 URLs would cost in the range from 28 000 – 75 000 Euro. This is beyond the current budget for the EIAO project, however, even if EIAO had the means to buy access to these data, it is not likely that it would be legal to redistribute NACE data extracted from the EBR.

2.1.1 National business registers

Of the 19 European Business Registers, 4 have been identified to provide NACE categorisation and address information freely via the Internet²:

Country	Business register	URL
Norway	Brønnøysundregisteret ([5])	http://www.brreg.no
UK	Companies house ([6])	http://www.companieshouse.gov.uk
Latvia	Lursoft ([7])	http://www.lursoft.lv
Serbia	The Serbian Business Registers Agency (SBRA - [8])	http://www.apr.sr.gov.yu

Most business registers charge money for searches, and only give out minimal information about the businesses for free.

Example of a national business register under EBR:

² Note that even if the information is freely available, there are terms and conditions that would prohibit automatic harvesting of this information by EIAO. There may be a few more, but not a large percentage. Furthermore, the search may have been limited by language barriers.

The Norwegian register for public and private businesses is Brønnøysundregistrene.

Basic information about Norwegian businesses publicly available via Internet via: <http://w2.brreg.no/enhet/sok/>

For example search for Universitetet i Agder (University of Agder):

<http://w2.brreg.no/enhet/sok/detalj.jsp?orgnr=970546200> gives:

Organisasjonsnummer:	970 546 200
Navn/foretaksnavn:	UNIVERSITETET I AGDER
Organisasjonsform:	Organisasjonsledd
Forretningsadresse:	Gimlemoen, 4604 KRISTIANSAND S
Kommune:	KRISTIANSAND S
Postadresse:	Serviceboks 422, 4604 KRISTIANSAND S
E-postadresse:	postmottak@uia.no
Internettadresse:	www.uia.no
Telefon:	38 14 10 00
Mobil:	-
Telefaks:	38 14 10 01
Registrert i Enhetsregisteret:	20.02.1995
Stiftelsesdato:	-
Daglig leder/ adm.direktør:	Tor Asbjørn Agedal
Kontaktperson:	Tor Asbjørn Agedal
Næringskode(r):	80.301 Undervisning ved universiteter
Sektorkode:	110 Stats- og trygdeforvaltningen
Også registrert i:	Merverdiavgiftsmanntallet, NAV Aa-registeret

This shows that UIA is registered with NACE code 80.301.

Even though searching in the Brønnøysundregistrene is free of charge, the complete list of businesses are not directly available. For example, in the web interface you can search for the name “Universitetet i Agder” and get all organisations related to this name. However, you cannot search for all organisations with NACE code 80.30 (Higher Education). NUTS location for the company may be inferred from the official postal address of the business register. The cost for access to such information would probably have to be negotiated with each national business register.

3 Automatic NACE / NUTS categorisation

In this chapter we present approaches for categorisation of web sites into NACE and NUTS categories. Note that we in this chapter only present automatic categorisation outside the field of machine learning. In chapter 4 we present approaches based on machine learning.

3.1 NACE and NUTS classifications based on fields in the Capgemini URL set

The NUTS code, may also be inferred from the Administration/Service Provider field in the Capgemini data list, and service name may to some extent be mapped to NACE code.

	Service	Administration / Service Provider	URL
(1)	Passports	FORT-DE-FRANCE	http://www.fortdefrance.fr
(2)	Income taxes	Service Public	http://vosdroits.service-public.fr/particuliers/N13.xhtml
(3)	Driver's licence	MONTELMAR	http://www.mairie-montelimar.fr
(4)	Car registration	Martinique	http://www.martinique.pref.gouv.fr
(5)	Building permission	MULHOUSE	http://www.ville-mulhouse.fr
(6)	Public libraries	MARSEILLE	http://www.mairie-marseille.fr
(7)	Certificates	PARIS	http://www.paris.fr
(8)	Enrolment in higher education	Université d'Auvergne	http://www.u-clermont1.fr

Table 1: Example of Capgemini data source

3.1.1 NACE categorisation

In Table 1, the services *Passports (1)*, *Income taxes (2)*, *Drivers licence (3)*, *Car registration (4)*, *Building permissions (5)* and *Certificates (7)* could be mapped to NACE 84.11 (General public administration activities). Furthermore, *Public libraries (6)* could be mapped to NACE 91.01 (Libraries and archives activities). For results on the accuracy / coverage of this see chapter 6.1.

3.1.2 NUTS categorisation

NUTS code could be estimated from a combination of the Administration/service provider field, limiting the search to the domain of the URL. For example, *Paris (7)* (France) could be mapped to NUTS FR101. For results of the accuracy / coverage of this see chapter 6.2.

3.2 Geographical lookup

NUTS is a geographical description. In this chapter we look at the possibility of categorising a web site into NUTS category based on the geographical information that is available or easily retrievable. Note that for EIAO, NUTS category is suppose to correspond to the location where the service is applicable, not e.g. the actual location of the web site.

As an example, the public web site if the UK Prime Minister Office, 10 Downing Street ([9]), has a clear reference to the specific location in London (Downing Street). Despite of this, the services UK Prime Minister is clearly applicable for the entire UK. Thus, the corresponding NUTS for this site should be for the entire UK, not only for London.

Similarly, the web site public Norwegian Governmental, government.no ([10]) is physically registered in the Norwegian capital Oslo. However, it is clear that the Norwegian government services is applicable for the entire Norway, not only Oslo. Thus the NUTS categorisation should reflect this.

3.2.1 Whois

Whois ([11]) contains information about a web site (domain)³ and usually an address, that is a candidate for automatically mapping domain to NUTS region codes, however the address format in Whois is not well standardised, so it is hard to do a reliable mapping. In some European countries, e.g. Denmark, address information is protected by the national data protection act, which means that address information is not available in Whois for these countries. In addition, there are license restrictions on any use other than personal use. This means that it is not straight forward to get access to Whois data from a legal perspective.

3.2.2 Hostip.info

Hostip.info ([12]) is a community-based project to geolocate IP addresses or domain names. The database is freely available. Providing an IP or domain gives you the physical location of the server.

3.2.3 NUTS Location from location hints in web site

Often location hints about where the web site is located may exist on some web pages of the web site. For example web pages with address information or other location information, since a public web site meant for a specific municipality often will present information relevant to that municipality. One option for NUTS location categorisation might therefore be to look for location hints while the observatory is crawling the web site and automatically categorise the web site to the NUTS code and NUTS level that seems to fit the identified location hints best.

4 Machine learning approach for NACE/NUTS categorisation

An automatic NUTS / NACE categorisation could be based on a machine learning approach.

³ Note that a web site might not be correspond entirely to a domain. Several domains may be scoped as one web site or one domain could be scoped into several sites. However, the most common situation is that a web site directly corresponds to a domain.

Furthermore, for NACE categorisation, the problem very much resembles the field of topic detection. A NACE category is business category. Naturally, a web site within a given NACE category is expected to have information within the topic of this field. For example, we can expect a web site with the NACE category 85.11 (Hospital activities) to actually contain information within the topic hospital activities. On the other hand, there is no obvious correlation between NUTS categories and topics.

Furthermore, the research area of topic detection which is a very well researched area in literature. In this chapter we treat the the problem as a classification problem.

For classification to be successfully applied in a field, each category needs to be distinguishable from each other. In our case, this means that the web sites within a given NUTS or NACE category needs to have certain properties that can identified and extracted.

Such properties are in the field of pattern recognition called features. Normally, whenever applying pattern recognition to a new field a large part of the work goes into determining what features represent the different classes well.

4.1 NACE categorisation using Nearest Neighbour based on the term frequencies

For topic detection, the features have traditionally been how often the actual words have been used (term frequency). [13]

As an example, we may expect an English web site within the NACE category 85.11 (Hospital activities) to have a higher probability of containing words such as *Hospital, Medical, Doctor*⁴ than a web site that is not in the NACE 85.11 category.

4.1.1 Term frequencies for NACE documents

A term frequency is the frequency of words in the document. I.e. the percentage of a word used within a set of documents. The most frequent word used in a set of documents will be the the term with the highest frequency and so on.

⁴ Note that such words are normally extracted from a training set, not defined manually.

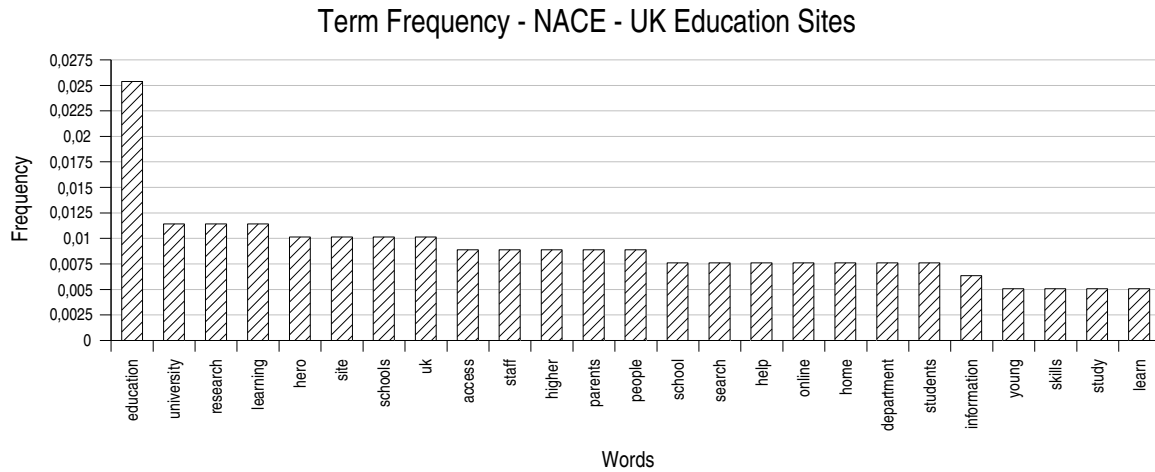


Figure 4.1: Term Frequency Education Sites

In Figure 4.1 and Figure 4.2 we show the 25 most common words and frequency of manually classified UK web sites as *NACE 80.00 Education* and *NACE 74.00 Investigation and security activities*, hereby referred to as Educational and Police.

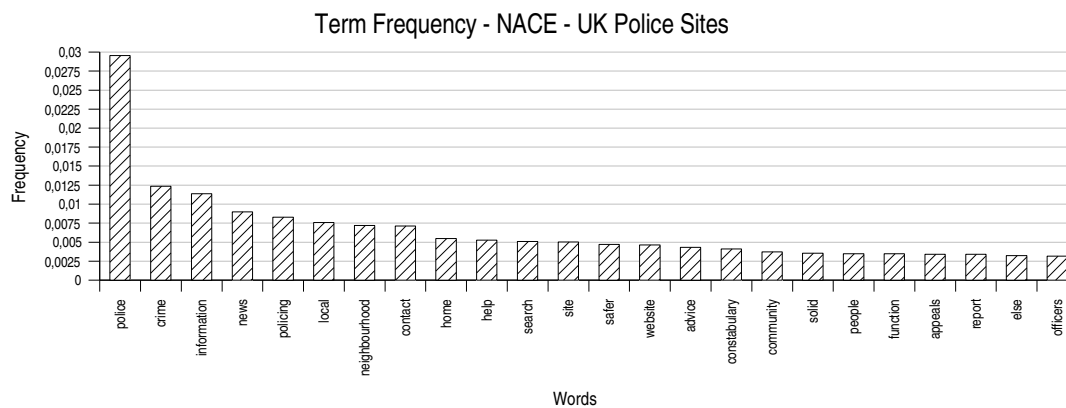


Figure 4.2: Term Frequency Police Sites

Note that these results have the stop words ([14]) removed, based on a predefined list of words, as is traditionally done in many text processing algorithms including topic detection. The stop words are typically small words such as: *and*, *to*, *for*. In our example, there is no reason to believe for example that the word *and* is more likely to appear in a police web site than an educational web site, and are because of this considered as noise.

By looking at the figures above, it is clear that there is a large difference between the web sites from educational and police web sites, which is a good indication that an algorithm should have the possibility of classifying such sites. In Figure 4.1 we see words that we typically identify with education such as university, learning. The same is true in Figure 4.2 where we see words such as police, crime, safer.

A much used algorithm for topic detection is Nearest Neighbour based on the term frequencies. For this algorithm, we initially train the classifier by determining the frequency order of the terms [15].

4.1.2 Training - NACE

As an example, for the education web sites shown in Figure 4.1, this order will be a vector of

$$v_{education} = [education, university, research, \dots]$$

similarly, for the police web sites in Figure 4.2 we get a vector of

$$v_{police} = [police, crime, information, \dots]$$

Note that the actual frequencies are not considered, only the order of the terms. I.e. for v_{police} the first term in the vector is police since this has had the highest frequency, the second term is crime since this had the second highest frequency.

This also ensures that both classes have an equal importance. An example of the opposite would have been to give police web sites a higher importance since we in this case have more training data for this class. However, in our case, we do not expect police web sites to be more common than education web sites, even though we have more training data. Because of this, we do not take the number of classes in the training into consideration. To ensure this we also need to cut the training vectors at the same point. In our example, the length of

$v_{education}$ and v_{police} needs to be the same as the maximum distance, used in the classification explanation in chapter 4.1.3 would be different if we did not.

The distribution of terms have been shown to be zipf [13]. Typically, only the middle part of such a distribution contains valuable information as the most frequent words are words used for classifying a language (such as stop words) and the least frequent words are considered noise.

4.1.3 Classification - NACE

Whenever a new web page i should be classified. A term frequency vector v_i is created based on the terms in i , similar to how we created $v_{education}$ and v_{police} previously.

Furthermore, the distance from v_i and each training vector ($v_{education}$ and v_{police}) is calculated. The distance defined as the sum of difference of each term in v_i compared to all training vectors. The class with the shortest distance to v_i , is the most probable class.

4.1.4 Proof of concept implementation - NACE

In this chapter we present an implemented for automatic NACE categorisation using the leave-one-out method. The leave-one-out method is defined as following:

For all web sites i

- Train the vectors, in this case v_{police} and $v_{education}$, for all web sites except i .
- Create the term frequency vector v_i .
- Classify v_i towards the vectors v_{police} and $v_{education}$.

Note that the important part of this approach is to not include a web site that is to be classified in the training vectors. In contrast, if v_i was included in the training sets the algorithm would be given wrong benefits as the classification data and training data overlap each other. Naturally, a functional classification algorithm should not expect the training data and classification data to be the same.

In Table 2 we show the a proof of concept implementation using the Nearest Neighbour tested with leave one out using UK Police and Educational web site.

		Correct classes		
		Police	Education	Accuracy
Suggested classes	Police	44	0	100%
	Education	0	4	100%
	Accuracy	100%	100%	100%

Table 2: Proof of concept implementation of NACE classification

It is clear from Table 2 that the over all accuracy is extremely high (100%).

A general rule in pattern classification is that when the number of classes increase, the accuracy of the algorithm decreases since the classification task becomes more difficult. In contrast, if the training set increases the accuracy of the algorithm increases as the training vectors become more complete.

Both of the above can be expected to happen in a real implementation of classification. In sum, we could expect the overall accuracy to decrease.

It is not so surprising that we get a perfect classifier with such small number of classes, specially when the terms are so different as seen in Figure 4.1 and Figure 4.2. In contrast, using all NACE categories we can expect a lower accuracy since the classification task is more challenging.

Further note that classification based on term frequencies is language dependant. For example, we cannot train the classifier on English web sites and expect that the classifier will work for web sites in other languages.

Also, several of the NACE categories are overlapping. We can for example expect some web sites to contain information both about Police and Education for a certain community. Thus, this web site will be included in both classes, which is impossible to classify in the above suggested solution. This is normally dealt with by adding a third category including both police and education. The training vectors would in our example then be

$$v_{police \setminus education}, \quad v_{education \setminus police} \quad \text{and} \quad v_{education \cap police}.$$

Never the less, this proof-of-concept of Nearest Neighbour based on the Term Frequencies indicates that such an approach is viable.

4.2 NUTS classification using a classification tree

Since NUTS is a graph structure, classifying web sites into NUTS categories using a graph algorithm, such as a classification tree, looks viable.

To take a simple example let us assume that we only have the following NUTS categories :

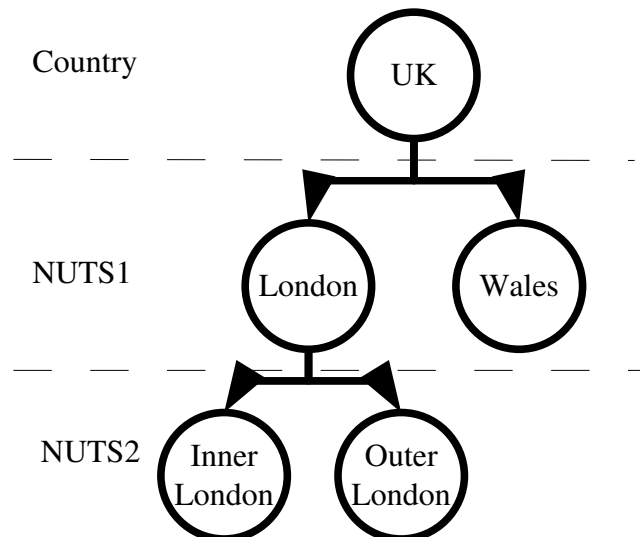


Figure 4.3: Example of NUTS categories

For NACE classification measuring accuracy is straight forward; it is clear that if an algorithm correctly classifies a web site, it is correct, otherwise it is wrong. For NUTS, on the other hand, measuring the accuracy is not that simple, which can be illustrated by the following example using example Figure 4.3

Assume that the correct class for a web site i is *Inner London*.

If the algorithm correctly classifies i as *Inner London*, the accuracy is 100%.

In contrast, if it classifies i as *Wales*, the accuracy is 0%.

However, if it classifies i as *Outer London*, the accuracy is not as wrong as *Wales*, however, it is not as correct as *Inner London*.

Thus, the calculation of accuracy for NUTS is not as straight forward as with NACE. This further reflects on the algorithm as the general approach is to maximize the accuracy.

The properties of NUTS, the tree structure of NUTS, very much resembles a classification tree [16]. In a decision tree, each node works as a possible class and each arc as a decision. One of the advantages is that the classification problem can be divided into several smaller problems. A traditional classification algorithm, not using a decision tree, would base itself on a set of classes and classify only according to these classes. In the example in Figure 4.3 we would expect the classes *UK*, *London*, *Wales*, *Inner London* and *Outer London*. However, using the example in Figure 4.3, we divide the classification task into tree smaller tasks.

An example of the process in such an algorithm would be, again assuming that the NUTS categories are limited to those presented in Figure 4.3:

1. Classify into Country.
2. If UK, classify into *London* or *Wales*.

3. If *London*, classify in to *Inner London* or *Outer London*.

4.2.1 Term frequencies - NUTS

We may expect a web site with the NUTS category UK1 (London) to have a higher probability of containing words such as *London*, *Westminster*, than web sites which are not part of this NUTS category.

In this chapter we present the term frequencies of 10 randomly selected UK and Ireland web sites.

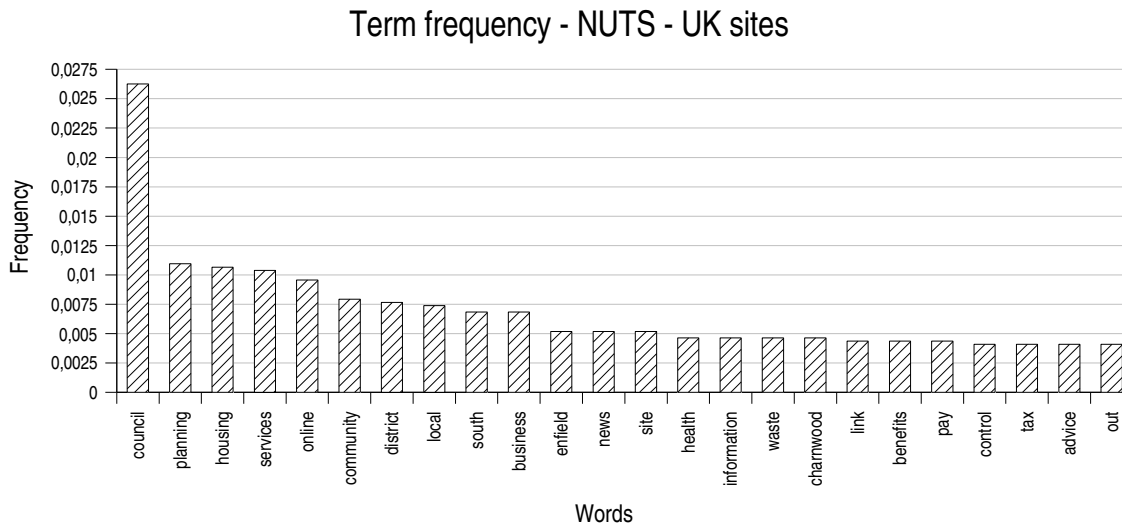


Figure 4.4: Term Frequency UK sites

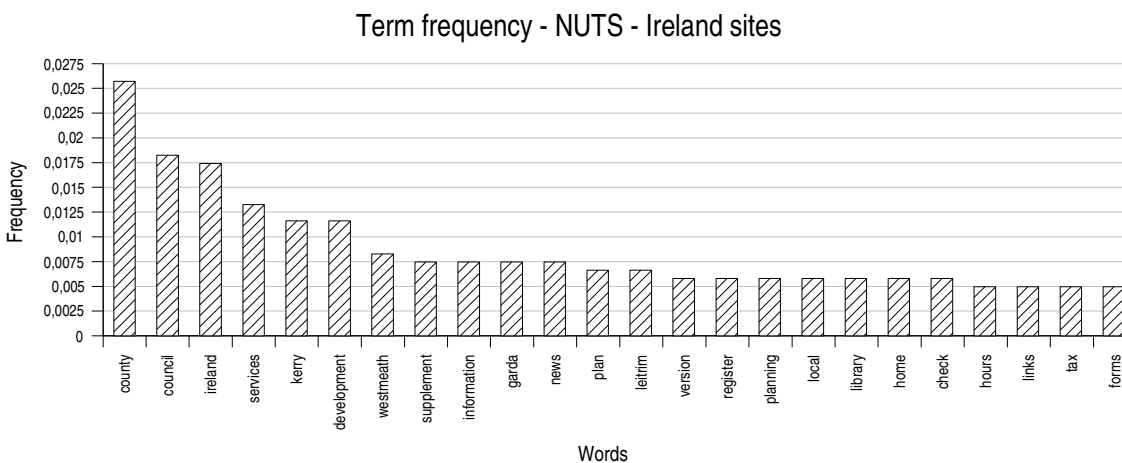


Figure 4.5: Term Frequency Ireland sites

In contrast to when the web sites were organised into NACE in chapter 4.1.1, we cannot see any clear difference in term frequencies from the UK web sites (Figure 4.4) and Ireland web sites (Figure 4.5). As an example, words such as *council*, *planning*, *services* and *local* are among the most common words for both UK and Ireland. However, some words are only visible in one of the categories such as *Ireland* and *Kerry* (for Ireland) and *distinct* and *business* (for UK).

The similarities in the data indicates, in contrast to NACE, that term frequencies are features that do not represent each of the classes well, which leads to a hard classification task.

Note that this approach automatically focuses on terms are frequent within a class. However, it might very well be that location terms are not frequent within these terms. From the above charts we can only identify Kerry and Ireland as location terms.

Note that we also here deliberately removed stop word ([14]) removed, based on a predefined list of words.

4.2.2 Decision tree - NUTS

Naturally, the decisions at each level needs to be defined in a sensible manner. One possibility would be to define the rules of decision prior to classification based on expert knowledge.

Assume that we plan to classify into NUTS2 (if we are already at level 3. in the algorithm outline).

Such rules could simply be as following; if the web site includes references to location in *Outer London* (e.g. *Harrow*), the decision should be *Outer London*. If the web site includes references to *Inner London* (e.g. *Camden*), the decision result should be *Inner London*. If the web site includes references to both classes (both *Inner London* and *Outer London*) or none of the classes, the decision could be NUTS1 – London.

The challenge with such rules is that defining these require expert knowledge of all NUTS locations. Creating rules are normally viable whenever the decision tree is small enough. However, for NUTS classification, assuming that we plan to classify into NUTS3, would require special rules for 2 492 categories – which clearly would require a substantial manual work. Note that many of such mappings already exist in the form of official location – NUTS mapping lists.

4.2.3 Classification tree - NUTS

Another possible option would be to treat each arc as an individual classification problem, thus creating a classification tree. As an example, classifying into NUTS2 (if are already at level 3. in algorithm outline) could be treated as a classification problem with the classes *London*, *Inner London* and *Outer London*. The decisions would then be automatically based on training data – for example based on Term Frequencies of the Nearest Neighbour.

This would result in as many individual classifiers as there is Countries and NUTS categories.

The advantage is that each problem would be relatively simple, and we can expect a high accuracy. The disadvantage is that, in order to properly train the classifiers, we need sample data of each class. As an absolute minimum, we need one web site for each possible NUTS class manually classified prior to the execution of the algorithm.

However, for classifying only into at most NUTS1, we would require number of countries + number of NUTS1 categories = 35 + 185 classes = 220 classes. Note that some of the countries have only one NUTS1 category, such as Norway. For these countries, classifying into country will be sufficient.

4.2.4 Proof of concept implementation - NUTS

Our proof of concept classifies web sites into countries: UK and Ireland.

Since UK and Ireland Web sites both are written in English, the language should not influence the accuracy of the classification algorithm. We believe that this classification resembles a classifier for two NUTS categories.

Note that, the proof of concept is only a Nearest Neighbour implementation of two classes using term frequencies, not a classification tree as suggested in chapter 4.2.3.

		Correct classes		
		<i>UK</i>	<i>Ireland</i>	<i>Accuracy</i>
Suggested classes	UK	10	0	100%
	Ireland	6	4	40%
	Accuracy	62.5%	100%	70%

Table 3: Proof of concept implementation of NUTS classification

As shown in Table 3 the over all accuracy is 70% using our small data set.

We believe that classifying into NUTS categories with a machine learning approach is a more challenging task than NACE categorisation. Our results indicate that for NUTS classification to be reliable we need to rely on other features than term frequencies, which is supported both by our findings in Figure 4.4 and Figure 4.5 and proof of concept implementation presented in Table 3.

This indicates that term frequencies alone are not the best approach for automatically classifying web sites into NUTS categories. Another possible option would be to, in contrast to using the term frequencies alone, that the feature may utilize a good location-NUTS database.

5 Manual Categorisation of NACE and NUTS

Manual categorisation of NACE and NUTS codes can base the categorisation on human knowledge about the business sector and location of the given web sites, by investigating the web sites. In this chapter we present an approach for manual categorisation. The accuracy of such a categorisation to some extent depends on the knowledge and experience of the human expert. For example, a person with knowledge of the German government can more easily and accurately categorise German web sites into both NACE and NUTS than a person without this knowledge. Because of this, it is preferable, whenever performing manual categorisation, that the expert has knowledge of the corresponding countries.

Note that we currently only categorise web sites into NUTS since this has been the highest priority.

5.1 Manual procedure for categorising a web site into NACE

- Use NACE revision 1.1.
- Use information from publicly available business registers, to learn how to categorise different services in the available URL set. Priorities where the business registers are freely available.
- To recognise a website of a city, municipality or similar, there is usually an emblem or a logo of the city.
- If the language is not understandable:
 - It is possible to search on Wikipedia ([17]) for the name of the website, or you can browse the website and look for an available picture (as an example there might be a picture of medical doctors when the web site is for a hospital). For universities, it is our experience that there are almost always English language version available, and the web sites include international terms such as Master, Bachelor, student.
 - Try to translate the whole page on “Babel Fish!” or similar ([18]). Such services can translate a web page into understandable language.
- The “Service” field in the Capgemini URL list maps a given service to NACE code relatively often.
- Some NACE categories are recurrent, like : commune: 75.11, university: 80.30, hospital: 85.11

5.2 Manual procedure for categorising a web site into NUTS

- Try to find the contact information of the web site.
- If the city is unknown, go to Wikipedia and search for it. Wikipedia usually provides information on country, state, and region. If there is no page about the city you are searching for, and you know what country it belongs to, you can try to search for the description of the country available at Wikipedia. There is usually a map which shows the region and the location of the city
- If the language is not understandable.
 - Try to find the English version of the website if it exists.
 - Try to translate the whole page on “Babel Fish!” or similar ([18]). Such services can translate a web page into understandable language.
- A large number of the links to services provided by Capgemini has already been categorised on a city level. Even though those not map directly to NUTS, it could be used as a basis for such a categorisation.

6 Comparing approaches

In this chapter we compare both the accuracy and applicability of the presented approaches for categorising web sites into NUTS and NACE categories.

6.1 NACE categorisation

We present preliminary results from experiments using manual categorisation and different automatic categorisation schemes with a selection of 225 URLs from the Capgemini set are summarised in the Table 4. Note that the machine learning approach has been applied to a different set of URLs than the remaining tests making the results not directly comparable. Further note that the accuracy has only been calculated for the machine learning approach. At last, the machine learning approach is designed to always provide a result thus always having a coverage of 100%. Further note that the comparison in Table 4 is not fair because the number of categories are different, which may result in the machine learning algorithm over performs.

Data source	Categorisation	Number Identified	Coverage	Accuracy
Machine learning	NACE	48	100%	100%
Manual categorisation	NACE	212	94%	-
Capgemini URL list, service column.	NACE	84	38%	-

Table 4: Accuracy and Coverage of NACE categorisation

In Table 5 we present a random comparison between manually categorised web sites into NACE with values registered in the Norwegian business registry ([5]). Based on this we can observe that the manual categorisation does not work very well. This is clear since the business register, which hold the authoritative information disagrees with the manual categorisation. Note that comparing automatic categorisation with public business registers has not been done.

Web Site	Manual Categorisation	Norwegian Business Register
http://norsk.lysingsblad.no/	92.51 NACE rev 1.1 Libraries and archives activities	22.13 NACE rev 1.1 Publishing of journals and periodicals
http://skien.kommune.no/	75.00 NACE rev 1.1 General public administration activities	75.13 NACE rev 1.1 Regulation of and contribution to more efficient operation of businesses.
http://www.sognefjord.net/politiet	74.40 NACE rev. 1.1 Advertising agencies	75.24 NACE rev 1.1 Public order and safety activities

Table 5: Random check of Manually categorised NACE code for some Norwegian sites.

6.2 NUTS categorisation

Preliminary results from experiments using manual categorisation and different automatic categorisation schemes with a selection of 225 URLs from available web sites set are summarised in the Table 6. The machine learning approach has been applied to a different set of web sites than the remaining tests and are thus not directly comparable.

We believe that the coverage and accuracy could be improved if the “Search for locations” solution included the entire set of pages crawled instead of only the home page, since location hints may be spread across the web sites and if the location mapping databases were improved to have increased coverage. Further note that the comparison in Table 6 is not fair because the number of categories are different, which may result in the machine learning algorithm over performs.

Data source	Categorisation	Number identified	Coverage	Accuracy
Machine learning	Country	-	100%	70%
Manual categorisation	NUTS3	214	95%	-
Search for locations on home page	NUTS3	57	26%	47%
Whois	NUTS3	44	20%	2%
Capgemini URL list, administration/service provider column.	NUTS3	40	18%	55%
Hostip.info lookup	NUTS3	14	6%	36%

Table 6: Accuracy and Coverage of NUTS categorisation

Furthermore, Table 6 shows the accuracy of each approach. Note that since the machine learning approach was applied to a different set of URLs attempting to categorise a web site into countries, rather than NUTS3, the

results are not directly comparable. Additionally, the machine learning approach is designed to always provide a result – thus the coverage will always be 100%.

Further note that for the remaining automatic schemes the accuracy is calculated percentages that matches the manual classification. In other words we assume that the manual classification has an accuracy 100%, which might not be true.

It is clear that manual categorisation achieves a far better coverage than most automatic categorisation schemes. Of the automatic schemes for NUTS3 categorisation, mapping from the Service description provided by Capgemini gives the best results.

To be able to correctly compare the results from all the schemes the machine learning approach needs to be adjusted to NUTS3 categorisation. Furthermore, the accuracy of manual classification needs to be calculated.

7 Conclusion

We have compared categorisation of NACE and NUTS using several approaches. Our experiments show that for NACE categorisation, a classification approach using term frequencies and the Nearest Neighbour (probably) is the most viable approach.

In contrast, automatic NUTS classification is a more challenging classification task for two main reasons; there exists a huge number NUTS categories and we have not been able to extract features that represent each of the classes well.

At this point, manual categorisation of web sites into NUTS categories is the most viable approach. Further viable possibilities for NUTS categorisation may include extending the location database and crawling the entire web site.

8 References

- [1] *The European Commission*, Nomenclature of territorial units for statistics - NUTS
Online:http://ec.europa.eu/eurostat/ramon/nuts/home_regions_en.html
- [2] *The European Commission*, Nomenclature Générale des Activités économiques dans les Communautés Européennes - NACE
Online:http://ec.europa.eu/comm/competition/mergers/cases/index/nace_all.htmlhttp://ec.europa.eu/comm/competition/mergers/cases/index/nace_all.html
- [3] *Capgemini*, Results of Capgemini's 7th Annual EU Online Services Study Released 2007
Online:http://www.capgemini.com/m/en/n/pdf_Results_of_Capgemini___s_7th_Annual_EU_Online_Services_Study_Released_.pdf
- [4] *EBR EEIG*, European Business Register Online:<http://www.ebr.org/>
- [5] *Brønnøysundregistrene*, Registerat og datakilde Online:<http://www.brreg.no/english/>
- [6] *Companies House*, Companies House Online:<http://www.companieshouse.gov.uk/>
- [7] *LURSOFT*, LURSOFT Online:<http://www.lursoft.lv/?&v=en>
- [8] *The Serbian Business Registers Agency*, The Serbian Business Registers Agency
Online:<http://www2.apr.sr.gov.yu/Default.aspx?alias=www2.apr.sr.gov.yu/eng>
- [9] *10 Downing Street*, 10 Downing Street - the historic office and home of the British Prime Minister
Online:<http://www.number-10.gov.uk/output/Page1.asp>
- [10] *Government Administration Services*, government.no, Information from Government and Ministries
Online:<http://www.regjeringen.no/en.html?id=4>
- [11] *Verio An NTT Communication Company*, Whois.net Domain-based Research Service
Online:<http://www.whois.net/>

- [12] *hostip*, My IP Address Lookup and GeoTargeting Community Geotarget IP Project
Online:<http://www.hostip.info/>
- [13] *Results of Capgemini's 7th Annual EU Online Services Study Released*, Results of Capgemini's 7th Annual EU Online Services Study Released
Online:<http://www2.sims.berkeley.edu/courses/is202/f05/LectureNotes/202-20051103.pdf>
- [14] *Wikipedia*, Stop words 2008
Online:http://en.wikipedia.org/w/index.php?title=Stop_words&oldid=209253980
- [15] *Toni Giorgino*, Results of Capgemini's 7th Annual EU Online Services Study Released
Online:<http://www.labmedinfo.org/download/lmi263.pdf>
- [16] *Wikipedia*, Decision tree learning
Online:http://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=196004793
- [17] *Wikipedia*, Wikipedia - the free encyclopedia that anyone can edit Online:<http://wikipedia.org/>
- [18] *Yahoo*, Yahoo! Babel Fish! Online:<http://babelfish.yahoo.com/?fr=avbbf-us>