Applied Text Analytics for Comments on News-Articles

A Bachelor Thesis

Anne Schuth
aschuth@science.uva.nl

ABSTRACT

Several on-line daily newspapers offer readers the opportunity to directly comment on articles. In the Netherlands this feature is used quite often and the quality (grammatically and content-wise) is surprisingly high. The paper develops techniques to collect, store, enrich and analyze these comments. After giving a high-level overview of the Dutch 'commentosphere' we zoom in on extracting the discussion structure found in flat comment threads; people not only comment on the news article, they also heavily comment on other comments, resembling discussion fora. We show how techniques from information retrieval, natural language processing and machine learning can be used to extract the 'reacts-on' relation between comments with remarkably high precision and recall.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Information Storage and Retrieval—*Content Analysis and Indexing*; H.4.3 [**Information Systems**]: Information Systems Applications—*Communications Applications*

General Terms

Algorithms, Experimentation

Keywords

web mining, web data extraction

1. INTRODUCTION

In the past few years, the World Wide Web saw some fundamental changes of which the phenomenon known as *user generated content* is the most striking; users, as opposed to owners, of websites are now able add their content. Probably the most extensively used form of user generated content is the blog, but recently also several daily newspapers offered their readers the possibility to comment on articles. These comments include much, high quality, discussion and debate, and could therefore potentially add substantial to the value and appeal of the on-line version of news articles. The liveliness, for example, in a discussion following an article might indicate how *hot* a subject is. Also, the way people

react on each other might reflect the social debate that is going on about subjects similar to that of an article.

For these reasons, and probably more, it is interesting to gain insight in the 'commentosphere'. This paper aims at bringing this insight in a way similar to what was done by G. Mishne [1], albeit on a much smaller scale.

Furthermore, on all of the news-sites it is presently not possible, for someone posting a comment, to specify whether he or she is commenting on the article or on some previous comment by another author. The effect is that discussions are often quite confusing and hard to read. It would be very helpful to bring forth the underlying structure, and to present this structure within the comment thread by, for example, links to posts the comment reacts on.

So, besides being of interest to social scientists, investigating the 'commentosphere' could also lead to enrichment of data already present, and new applications could be based on this additional data.

1.1 Research Questions

As mentioned above, the result of analytics used, and methods developed in this paper, could be of much interest to social scientist. Therefore the research question guiding this paper falls apart in two questions. One aims to provide insight in characteristics of the 'commentosphere', the other focuses on developing methods to come to those insights.

1. The first question, of interest to social scientists is: *What does the 'commentosphere' consist of and how does it look?* This is an umbrella for questions like: Who is posting comments? Where do commenters come from? Do they react on each other? In what way do they react? How does a comment look? What kind of language is used?

2. More of interest to information science, is the second research question: *How can properties of the 'commentosphere' be brought to light?* Again, this question is broad, and answering it would involve at least answering the following questions: How can text analytics methods be applied to comments? How can a person behind a comment be identified? How can the structure of comment threads be made explicit?

1.2 Organization of the paper

We first describe the data 'as-it-is' and how it is acquired, in Section 2. We then follow with applying some text analytics to the data, in Section 3, giving us partial answers to both research questions. Then Section 4 goes on with

develop ping methods to extract the structure of threads. Section 5, on Future Work, elaborates on how findings of this paper could be of further use and we conclude with conclusions in Section 6.

2. DATA DESCRIPTION AND ACQUISITION

In this paper mainly two different types of data are considered. These are content, in the form of news-articles, generated by a professional author for a news-site. Generally it can be said that the text they generate is checked on spelling, grammar and punctuation errors. As opposed to the other type of data, the *user generated content* (UGC), posted as comments, by commenters, on the news-articles mentioned before. These are likely to be full of mistyped words and made-up words, they might lack punctuation and it is not guaranteed that there is any grammar at all, although at a first glance it seems that comments, in fact, are of a notably high quality.

2.1 Concepts

In the above, four concepts came up: site, article, comment and commenter, below a more precise definition of these concepts that will be used throughout this paper.

2.1.1 News-site

The paper focuses on dutch news-articles and the comments on these only. Some limits where needed and since the author is dutch and therefore familiar with the dutch newspapers, this seemed a reasonable choice.

In the Netherlands, there are few nation wide pressed newspapers, namely: *Telegraaf, Trouw, Algemeen Dagblad, NRC Handelsblad, Volkskrant, Financiele Dagblad*. All of these have a website version with at least the current articles readable, but only a few have the option to leave a comment on an article. Since the interest of this paper is comments on news-articles, only those site with the possibility to comment are used. Additionally the site http://www.nu.nl is added. This is web-only news-source that might serve as valuable comparison material. From now on, the word *site* will indicate a source that produces news-articles. Below the chosen sites with some remarks.

Trouw. has a website on http://www.trouw.nl where only a few articles have the comment option. Besides that, very few articles are posted every day.

Algemeen Dagblad. hosts a website on http://www.ad.nl with many new articles every day and with the option to comment on all its articles.

Telegraaf. on http://www.telegraaf.nl, is similar to Algemeen Dagblad with respect to the amount of articles and the option to comment on all of them.

Nu.nl. is a little different. The website itself does not have any comment option. Instead, is has a link to http://www.nujij.nl which is a site that only hosts comments (no articles) on various websites. For this paper, only those article from nu.nl are used that have at least one comment on nujij.nl

NRC. again, is a different story. It is not possible to comment on all articles, instead there are a few discussion articles, created especially for that purpose, one every few days, that do support comment. This site is added because a more serious discussion might be expected, and therefore it might serve as an interesting comparison.

2.1.2 Article

An article is the content published on a news-site. It might have the option for users to comment on it. Or it might not, as it is with most articles on the Trouw and NRC website. Usually the author who wrote the article is mentioned, as is its place of origin. All news-sites also state the date and time of publishing, except for Trouw, who states only the publishing dates, and no times, this will be a shortcoming as will be seen later on.

2.1.3 Comment

The comments on articles consist of at least the name of the author and the comment itself. At a few sites, an email address and the city the commenter lives in, are also required. As is the case with articles, comments have, on all sites, a date and time of publishing attached to them, no exceptions here. Although, nujij.nl, the source of comments on Nu.nl articles, displays the time of publishing in elapsed time. Meaning that each comment is accompanied with a remark like: *posted 3 hours and 34 minutes ago*. After one day the hour and minute notation is replaced by a *posted 1 day ago* statement. On some sites comments get edited by employees of that site due to flaming, discrimination in these comments.

2.1.4 Thread

For those articles that have comments attached to them, we speak of a comment-thread, or just thread. It is a flat list, linear in time consisting of comments. There are no explicit relations between comments other than the publish-time, as opposed to, for example, fora were usually features exist to mark a post as a reply to a specific previous post.

2.1.5 Commenter

A commenter is person who comments on an article. Usually a commenter only leaves a name behind, but on some site a city and email address as well. Commenters may comment on an article more than once, and they can comment on more than one article on more than one site. Since usually only a name indicates a commenter it is often quite hard to be certain about the identity of a commenter.

2.1.6 Category

All the five sites use different or no categorizations to categorize their articles. See Appendix A for an overview of the correspondence between the categories used on the sites and those used in this paper.

2.2 Acquisition

It is the purpose of this paper to investigate the properties and structure of comments on news-sites. In order to be able to do this, data needs to be acquired and put in a format that makes it easily accessible.

2.2.1 Retrieval

All five of the new-sites offer a RSS-feed with headlines and links to the newest articles. Only NRC publishes the full content of posts, that can be commented on, through

their feed. So for all sites but NRC the articles need to be extracted from the HTML version of the site. Specialized wrappers are written to fit each site an make sure as little HTML as possible is attached to an article.

Scraping comments is done separately from scraping the articles, the reason for this is a timing issue, discussed later. As the articles provide a location for the comments (usually the same link, or in the case of nu.nl a link to nujij.nl), these locations get scraped completely, no matter how many comments are in the database already. The total amount of scraped comments is then compare to the amount available in the database and any surplus comments are added to the database, avoiding unnecessary data flow and overwriting possible edits by editors of the sites. The harvesting of comments is, as was the case with most articles, done through the HTML interface of the news sites, with specialized wrappers for each site.
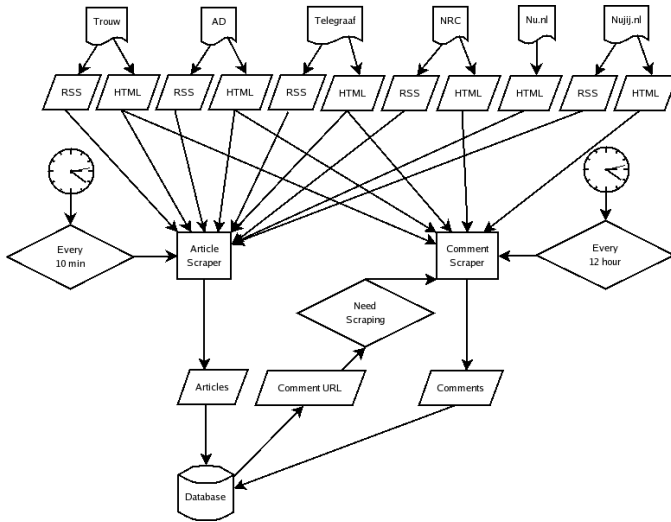
The complete data-flow is illustrated in Figure 1.

Figure 1: Flowchart of the data acquisition process.

2.2.2 Storage

All the collected data is simply stored in a relational database, mainly consisting of three relations: *site*, *article* and *comment* with attributes and their type as listed in Table 1.

site	article	comment
id *(int, key)*	id *(int, key)*	id *(int, key)*
name *(string)*	site_id *(int, foreign)*	article_id *(int, foreign)*
url *(string)*	source_url *(string)*	full_text *(blob)*
	comment_url *(string)*	author *(string)*
	title *(string)*	city *(string)*
	description *(blob)*	email *(string)*
	full_text *(blob)*	date_published *(time)*
	date_published *(time)*	date_scraped *(time)*
	date_scraped *(time)*	

Table 1: Relation attributes

2.2.3 Character encoding

It turned out that character encodings would deliver quite a serious problem to the automated collection of data. Dif-

ferent sites use different encodings. And commenters sometimes copy-paste something into their comments that leaves behind traces of the source encoding of this their texts. The result is a whole bunch of different encodings that need to be dealt with. The easiest solution to this problem, that covered most cases, was converting everything to *utf-8* encoding. Still some issues remained resulting in errors in the scrapers, however, these errors only occurred in a very marginal portion of the cases and is therefore not dealt with. Also, in different encodings different quotations are used, this is resolved by replacing all quote-like characters by their HTML-escape variant, avoiding trouble in SQL commands.

2.2.4 Timing

The time of acquiring data should be in line with the publishing times of data. After some experiments and observations, it turned out that a ten minute interval in between two sweeps along all five of the sites made (more than) sure that no new articles in the RSS-feeds would be missed due to a limited feed-length.

The collection of comments should not be done immediately after the article was published, since it will take some time before commenters will write their comments. And since there are no RSS-feeds for the comments, there is no limit to the number of comments that can be retrieved. Scraping twice a day ensures that the database is reasonable up-to-date all the time. Besides that, scraping twice a day will also ensure that a comment will always be scraped within 24 hours which has the advantage that the hours and minutes of the publishing time of nu.nl comments get preserved, see Section 2.1.3.

2.2.4.1 Duration.

The scraping of comments takes a considerable amount of time, and this is another reason to only scrape twice a day. To give an indication, on a Intel Celeron CPU 2.8 GHz the time t in processing minutes is given with:

$$t \approx 1.5 \cdot 10^{-2} \cdot \alpha$$

Where α is the number of articles, and *not* the number of articles with comments, since this is what is discovered during the scraping. This duration, off course, varies with the CPU type, Internet connection speed, number of comments per article, the distribution of articles over the sites and the implementation of the scraper.

Since the time it takes to collect all comments goes up by the number of articles, and scraping needs to be done within 12 hours, it makes sense to set a moving time-frame for collecting comments. Figure 2, later described in more detail, shows us a histogram of the response time relative to the publishing time of an article. It can be seen that the tail of the different curves differ greatly; different sites attract comments longer or less long after the publishing of an article. This gives us an indication to when to stop harvesting comments. A good heuristic would be to harvest until a certain percentage of all comments, per site, are in, based on what can be estimated from precious results. This should be done in such a way that the time it takes does not exceed 12 hours, $t < 60 \cdot 12$. Taking for example 95% as the desired mass of comments, would give the time frames as in Table 2[1].

---

[1]The time frame for Trouw is not very acurate since pub-

site	frame
Trouw	0-23 hours
AD	0-31 hours
Telegraaf	0-29 hours
Nu.nl	0-46 hours
NRC	0-332 hours

Table 2: Time frames for checking harvesting comments after an articles is published.

3. DATA EXPLORATION

In order to gain some insight in the data considered and to give ground to choices made later on, some very basic statistics followed by a more in depth investigation of some of the phenomena encountered is given in this section.

3.1 Statistics

3.1.1 Articles

As explained earlier, articles are those text published on news sites by professional authors. Table 3 shows the mass of the data acquired by the acquisition process[2], Section 2.2.

site	days	articles	
		total	per day
Trouw	47	285	6
AD	35	8139	232
Telegraaf	42	7198	171
Nu.nl	85	2109	24
NRC	71	20	0
		17751	

Table 3: The amounts of articles, per site. Trouw has very few articles available, as has NRC. Both AD and Telegraaf publish vast amounts. For Nu.nl only articles with comments on nujij.nl are indexed.

The number of days that are included in the data set differ per site due to differences in feed lengths, and in the case of Nu.nl due to the fact commenters may react on a article days or even months after it was published, causing it to show up in the nujij.nl feed. Trouw and NRC both show very low numbers of articles, compared to the other sites, this is due to the fact that both allow only a few articles to be commented on. Also, the numbers for Nu.nl are a little lower than those of AD and Telegraaf because only those articles with comments on nujij.nl are indexed.

3.1.2 Comment threads

A comment thread is a collection of comments all belonging to one article. First of all the amount of threads is shown in Table 4.

Even though for NRC and Trouw only those articles that allow for commenting are indexed, the percentage of the articles that have a thread attached to them are not 100%. For NRC this can be acounted to the deletion of articles from the website, making it impossible to comment on them. For Trouw it can simply be acounted to the lack of people

---

lishing times of articles are not always given, it is probably wise to expand this frame, even more because Trouw does not seem to attract much comments.

[2]These figures were accurate on the morning of the 26th of June 2007

site	articles with comments		number of comments	thread-size		
	total	%		avg	min	max
Trouw	80	28%	1297	16	1	88
AD	3465	42%	33383	9	1	202
Telegraaf	636	8%	45760	71	1	826
Nu.nl	2088	99%	34165	16	1	689
NRC	19	95%	2557	134	34	314
	3431	35%	117162	23		

Table 4: Some meta data about comment threads, split up per site.

commenting on an article, while it is possible. For Nu.nl also a 100% would be expected, but although it is close to 100%, also there some thread got deleted. AD and Telegraaf show over time a rather stable percentage of articles with a thread attached to them.

Then, if comments are published, it is interesting to see when, as in how long after an article was published, this is done. And for how long commenters keep publishing their comments. Figure 2 gives an impression of the amounts of comments that react on a article for the first week after an article was published.
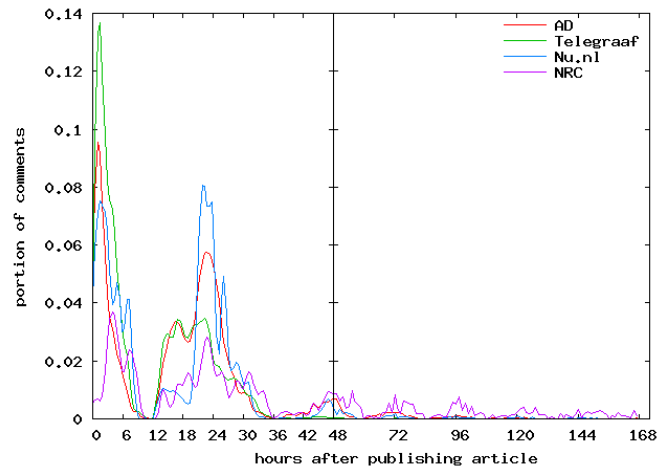

Figure 2: The average timing of publishing comments after an article is published, for a week. Note that the data for Trouw is left out because no publishing time is provided. Also note the differences in scaling on the x-axis, used to give a more detailed view of the first 48 hours.

The figure shows the time, in hours, of posting a comment plotted against the portion of comments that was posted at this delay; this gives effectively a histogram. Data for Trouw is left out because Trouw did not always provide for the time of publishing an article, sometimes only the date was given. One might think that the time of scraping provides a clue to the time of publishing. However, this is only true for those articles that appeared in the RSS feed after the scraping process was started, not for those that were in the feed already.

The graph in Figure 2 clearly shows a high peak right after the article was published, and then declining peaks, each 24

hours later. These peaks are *not* the result of the scraping that is done twice a day, Section 2.2.4. They are the result of much less people commenting during the night. It should be noted that the width of a peak is partly a result of the spreading of publishing times of articles during the day.

Also note that peaks for the NRC site are much lower than the others, but they keep recurring for a much longer time; on the NRC site, commenters keep commenting on an article as much as a week after it was published. This as opposed to the curve for the Telegraaf that shows almost no new comments after as little as 36 hours after the article was published, which is not the effect of closing the comment thread.

3.1.3 Comments

We now look at some of the properties of comments itself. A tokenizer, as described in [4], was used to split up comments in words and sentences. First, Figure 3 gives the number of sentences per comment as a histogram; the number of sentences used is plotted against the number portion of comments that haves this number of sentences. Portions, instead of amounts, are used to make comparing sites possible.
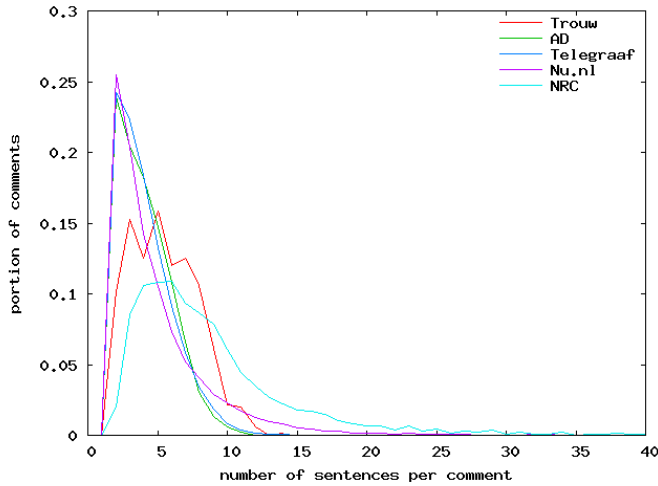

Figure 3: **A histogram of the number of sentences per comment. Each point on a line indicates the portion of comments that has the number of sentences measured on the x-axis.**

There seem to be two types of curves here, the very steep and high ones around three sentences per post, for Telegraaf, AD and Nu.nl, and the much less steep ones having a peak around 6 sentences per post, for Trouw and NRC.

Another interesting property of comments is the length of sentences used in post. Figure 4 plots the average sentence length per comment against the portion of comments that have this property. As was the case with the number of sentences, Trouw and NRC jump out again. The average sentence length in comments on articles for those two sites is around 13 words per sentence, while the other three sites have comments with an average sentence length of around 9 words.

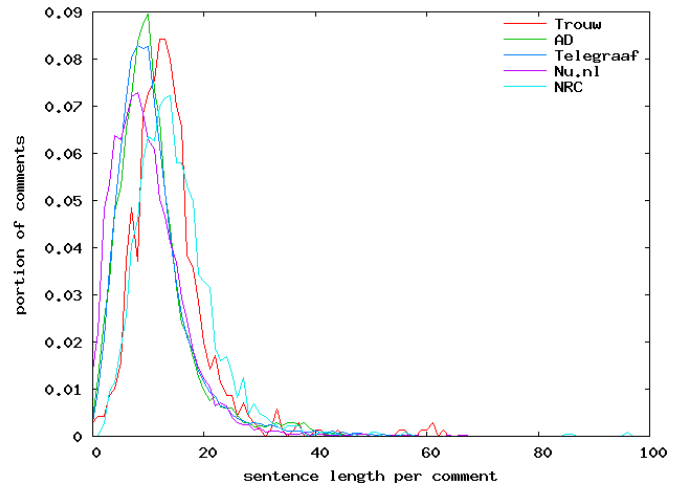Both the number of sentences and the average length of

Figure 4: **A histogram of the average length of sentences, in words, per comment. The sentence length is averaged per comment and the size of the portion of comments having the same average sentence length is plotted against the sentence length.**

these sentence is significantly higher for comments on the NRC and Trouw site than they are on the other three sites. Evidently, this means that comments on those two site are much more substantial.

A very different property of comments can be the sue of special characters. A manual investigation showed that the @-character was used quite frequently to indicate a reference to another commenter. This was the reason to count the use of this character in comments, as it might give us a first glance at the relations that might exist between comments. Table 5 shows an overview of the difference of the usage of the @-character on sites.

@count	Trouw	AD	Telegraaf	Nu.nl	NRC
0	1.0	0.8607	0.9889	0.7944	0.9707
1	0	0.1326	0.0108	0.1925	0.0269
2	0	0.0061	0.0003	0.0107	0.0013
3	0	0.0004	0	0.0018	0.0004
4	0	0.0002	0	0.0003	0
5	0	0	0	0	0

Table 5: **The @count column indicates the number of @-characters found in a comment. The other columns show the portion of comments that had this many @-characters in their full-text attribute, split up per site. As can be seen, no more then four @-characters were used in any comment.**

These number show that on the AD and Nu.nl sites, there is substantial use of the @-character while it is very marginal on the other sites.

3.1.4 Commenters

Another completely different type of data are the commenters. Of them, in most cases only a name is known. And on some sites also the city they live in and their email address is given. As the names commenters use, are chosen by themselves, it is hard, if not impossible, to distinguish

between two commenters that call themselves, for example, *John.* Other, more exotic names, seem to provide more evidence that the name is indeed referring to just one person but a 100% quaranty is never given. The cases were one can be sure, are those were an email adress is provided, since these systems all work with authentication emails.

Even though the issues above are serious, it does give insight in the data when comments are grouped by email addresses and commenter names. The graph in Figure 5 does this, and plots each commenter against the number of comments he posted.
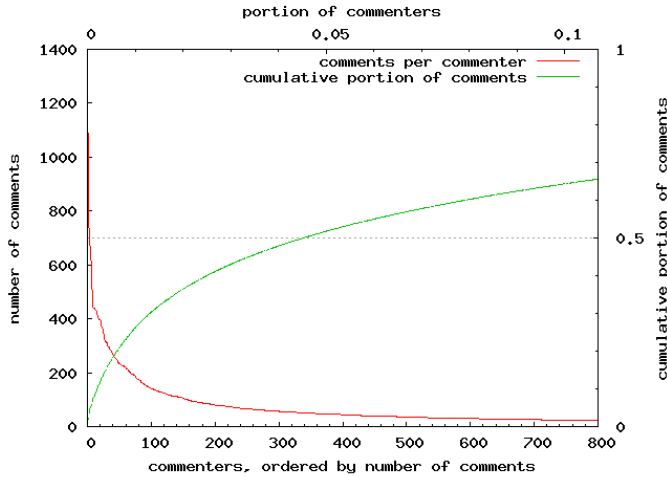

Figure 5: Commenters' names are grouped by email address and name and are then plotted against the number of comments they posted. The top and bottom axis correspondent, the left and right axis do not. The left axis belongs to the red curve, the right to the green. Less than 5% of the commenters post 50% of the comments.

The graph shows some very interesting properties; 50% of the comments is posted by less than 5% of the commenters. Ofcourse, we should discount for the very common names, of which many apear in the top 100 of most posting commenters. How this should be done exactly is not clear, but even when these names are filtered out, it is evidently that a large portion of the comments is produced by a relative small number of people. The distribution seen in Figure 5 is called a Power-law distribution and is observed in many domains [?]. The distribution is very common in social networks. To give an impression on the names of commenters that posted a vast amount, Table 6 gives the top 10, with common names filtered out by hand.

As some sites allowed for entering the city a commenter lives in, it is interesting to see what the geographical distribution of commenters is. Using the geocoder from Google Maps[3], all coordinates for cities given in by commenters were requested. Those cities which coordinates in between $(50°\text{N } 3°\text{W})$ and $(54°\text{N } 8°\text{W})$, roughly the coordinates of a bounding box of the Netherlands, were selected and plotted in Figure 6.

---

[3]see: *http://www.google.com/apis/maps/documentation/ index.html#Geocoding_HTTP_Request*

Name	Number of comments
info	621
G. Henkie	509
ef.jonker	500
Eb	444
firdausi	431
Steinmann	393
Jozias	367
Harrie_H	359
ooievaar	314
EET GEEN	313

Table 6: The top 10 of commenters with their post count. Left out are the very common dutch names like: Jan, Peter, Rob, Hans, Henk, Johan, John, Kees, Jeroen, Daniella, Piet, Erik, Marco, Paul, etc.
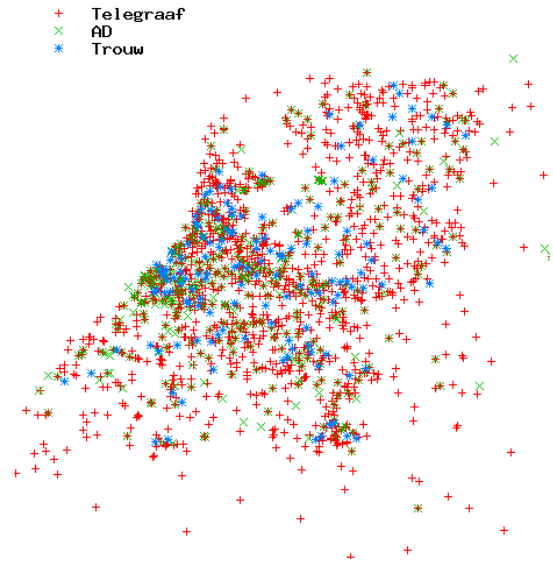

Figure 6: The geographical location of commenters, between $(50°\text{N } 3°\text{W})$ and $(54°\text{N } 8°\text{W})$, plotted for the sites that support this. Trouw data is plotted on top of AD data, which is plotted on top of Telegraaf data since the Telegraaf, followed by AD, had most data available. The contours of and the most densely populated areas in the Netherlands can easily be recognized.

As can be seen, this plot resulted in quite a detailed map of the Netherlands with a high density of commenters in the so-called *Randstad*, which is the most highly populated area.

4. THREAD STRUCTURE

It is expected that comment threads have some sort of structure; even a first glance at the data shows that commenter react on each other. However, as was mentioned earlier, this structure is not explicit; commenters do not have the option to comment on a specific comment, as is the case on for example most fora on the web. A big part of this paper is about unraveling this structure. The structure we are looking for is a *rooted directed acyclic graph*, or a *rooted DAG*. The root is the article, and the directed and acyclic properties are enforced by time. The nodes in this graph are comments and the arcs are a *reacts-on* relation. The struc-

ture is not a tree since it possible for one comment to refer to more than one comment, and for more than one comment it is possible to refer to the same comment.

4.1 The reacts-on relation

As it is used above, and in the rest of this paper, by stating *comment A refers to comment B* means that there is a *reacts-on relation* from A to B. By *reacts-on relation*, we mean that there is a relation between two comments as a result of one comment reacting on the other, a formal definition follows. The reacts-on relation is never reflexive: $\forall a \in Comments, \neg reactson(a, a)$; a comment can not react on itself. It is also never symmetric: $\forall a, b \in Comments, reactson(a, b) \Rightarrow \neg reactson(b, a)$; a comment can not react on a future comment. Also, the relation is not necessarily transitive: $\forall a, b, c \in Comments, \neg(reactson(a, b) \land reactson(b, c) \Rightarrow reactson(a, c))$; if a comment a reacts on a comment b, and this comment b reacts on a previous comment c, it is not necessarily the case that comment a also reacts on comment c, although this *could* be the case.

4.2 An investigation

In order to get a first impression on the discourse structures within threads, a random pick of 5 threads was made out of a total of the 2875 threads with 2 comments or more, that were available at that time. The 5 threads are listed in Appendix B, as a reference.

Only 12 *reacts-on* relations were found between the, in total, 50 comments. Most of these were in the longest thread of 24 comments. A manual investigation leads to the following observations with respect to the *reacts-on* relation:

1. Referring happens by citing an authors' name ('Eens met "bringer of torture"')

2. A referring statement might be misspelled (*Frabkster* instead of *Frankster*)

3. One comment can refer to more than one author ('Bravo *Djieff* en *Frabkster*')

4. One comment can be referred to by more then one other author (comment 2 and 4 on *Markplaats bereikt mijlpaal*, both refer to comment 1)

5. The '@' character may be used to indicate a reference. (*@ouwerocker*)

6. A citation usually seems to refer to the latest post by the cited author (the next observation might be an exception)

7. It is not always clear to which post of a cited author is the citation is referring (e.g.: *J.J.*, in comment 11, might refer to comment 8 or 7)

8. Quoting a part of a previous post, that includes a citation does not necessarily lead to a new reference.

9. Referring to a comment does not necessarily involve citing a name, a reference may simply state that is does not agree with something posed by another author, thereby very implicitly making a reference. (see comment 2 on *Markplaats bereikt mijlpaal*, it refers to comment 1. Or comment 4 in the same thread, it refers to comment 2)

10. The type of *reacts-on* relations differ:

   - **Agreement**: comment A agrees with comment B
   - **Disagreement**: comment A disagrees with comment B
   - **Advice**: comment A gives advice to the author of comment B
   - **Answer**: comment A answers a question in comment B
   - **Discussion**: commenters A and B keep reacting on each other

4.3 Limitations on the reacts-on relation

Section 4.2 lists some observations all relating to the *reacts-on* relation. These observations could be used as an inspiration for devising a method that would find the structure in threads. The thread structure would be quite complete once this method would be able to correctly identify all 10 observations. It is however not feasible, at least not by using simple, mostly syntax based text analytics, as was done in this paper to achieve this goal. The main reason for this is that some issues, observation 7 through 10, require deep semantics to be analyzed. Therefore we will only attempt to find reacts-on relations of the type that comply with observation 1 through 6; in other words, we will only attempt to find relations that involve citing an authors name. We do, however, allow for misspellings in the referring statement and multiple relations from and to a comment.

4.4 Measuring progress

Since there is need for a way to measure the quality of a method and the progress we are making, it seems like a good idea to have the most basic and simple approach, that will do reasonable but not yet *good*, to serve as a baseline.

4.4.1 Evaluating

To evaluate the performance of a method, the found relations could be compared with hand-annotated data. This comparison will tell the number of correct (*true positives*) and wrong (*false positives*) relations, which can be expressed as recall, and respectively, precision. The general formula for recall is:

$$recall = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

In our case, this translates to:

$$recall = \frac{|\{\text{correct relations}\}|}{|\{\text{annotated relations}\}|}$$

Where $|\{\text{correct relations}\}|$ indicates the number of correct relations returned by the method under evaluation, and $|\{\text{annotated relations}\}|$ is the number of relations that *should*, in the ideal case, have been found. The general formula for precision is:

$$precision = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

This, again, translates to:

$$precision = \frac{|\{correct\ relations\}|}{|\{incorrect\ relations\}| \cap |\{correct\ relations\}|}$$

Where $|\{incorrect\ relations\}|$ indicates the number of incorrect relations returned by the method under evaluation. Besides these two measures, we make use of the F_α-measure:

$$F_\alpha = \frac{(1 + \alpha) \cdot (precision \cdot recall)}{\alpha \cdot precision + recall}$$

The F_α-measure is the general case of the F_1 formula that weights recall and precision evenly. For different applications, a different α might be desired valuing precision and recall at different degrees. For the rest of this paper we make use of the F_1-measure. In Section 5 we speak about an application that might involve attaching more weight to precision.

The *false negatives* are simply $1-recall$ and the *true negatives* are $1-precision$, both measures do not give additional information and are therefore left out of the evaluation.

4.4.1.1 Annotated Set.

It is not feasible to annotate all data by hand, but a small portion, chosen semi-randomly that at the least includes a sample thread for each site, would give a fair enough indication of progress made. It is probably wise to make sure that each site has at least one thread in the sample, since it is not unthinkable that the commenters methods of referencing, differ from site to site (think of the use of @-characters, table 5). Even so, using rather large threads will more likely ensure that there is anything at all to evaluate.
Eventually the set of comments consisted of 518 comments in which a total of 193 reacts-on relations was identified, details of number per site can be found in Table 7. It should be noted that this set is completely separate from the set used in the first investigation in Section 4.2. This is done because here rather large threads are used that did not serve very good as examples to be attached in the appendix.

4.4.1.2 Over-fitting.

Another remark that should be made, when using this method for evaluation, is that it is very easy to over-fit the data chosen as sample set. This is because incrementally evaluating more sophisticated methods and using the evaluation results to tweak those methods even more will lead to a method that is very good at solving the problems at hand in the annotated set, although it very possible is not able to handle lots of phenomena encountered in the rest of the corpus. To overcome this shortcoming, a final evaluation, on which no further improvements are based, should be performed. This final evaluation should, again randomly, select some relations from the pool of relations found by the method in question. These selected relations should be evaluated by hand, leading to a final precision measure. It will be hard to evaluate recall this way.

4.4.2 Baseline

As said, the baseline should be a simple measurement. An obvious and simple approach to automatically find relations between comments is to match all preceding commenters names to a comment. And when a match is found,

interpreted this as finding a relation. Opting for this, a definition of *matching* is needed. Looking at the investigation in Section 4.2, a case-insensitive string matching, between each author and the full text of the comment, seems reasonable. A problem with this method is that it will find names included in normal words as well (i.e.: *Boy* in *lover-boy*). On the other hand, a more sophisticated method that would look at word boundaries, would have difficulties with names spanning more than one word (i.e.: *bringer of torture*). Clearly, these two issues could be solved but that is not the objective here. A starting point is needed from where progress can be measured. The baseline method is, on a high level illustrated in Algorithm 1.

Algorithm 1 Baseline
___

  initialize the set of Commenters as $\emptyset$
  order Thread ascending by time
  for each comment $\in$ Thread **do**
    for each commenter $\in$ Commenters **do**
      if commenter_of(comment) $\neq$ commenter **then**
        if in_string(comment, commenter) **then**
          assert reacts_on(last_post_of(commenter), comment)
        end if
      end if
    end for
    add commenter_of(comment) to Commenters
  end for

Using the method sketched above and evaluating as in Section 4.4.1, gives table 7, which format will serve as a standard for evaluating methods in this paper, and which data will serve as a baseline to compare more sophisticated methods to. In subsequent tables, similar to table 7, the number of comments and relations will be left out and two columns, with the difference (δ) to both recall and precision of the baseline, will be added

	comments	relations	recall	precision	F_1
Trouw	40	0	-	0.0	-
AD	34	9	0.8888	0.2580	0.4000
Teleg.	56	8	0.3750	0.1363	0.2000
Nu.nl	313	163	0.4723	0.5789	0.5202
NRC	75	13	0.6923	0.1764	0.2812
			0.6071	0.2874	0.3901

Table 7: Recall and precision for the baseline method. Also, the numbers of comments and annotated relations involved.

It is often, as it is here, easy to get a high recall, at the cost of a low precision, by simply returning all possible relations, and vice verse. The objective of the paper is therefore: maximizing both *recall* and *precision* at the same time.

As can be seen, Trouw has zero relations in the annotated thread. This is because no thread with relations could be identified by hand. This does not necessarily mean there is no such thing as commenters reacting on other commenters on the Trouw site. The methods below might in fact be able to find some, even though a manual investigation could not find any. Because Trouw has zero relations in the baseline it is hard to talk about improvement, therefore, Trouw is left out of most evaluations.

4.5 Advancing methods

As mentioned in Section 4.4.2, the baseline has some very obvious disadvantages some of which can be solved relatively easy. Some others involve some more sophisticated techniques. All of these will be discussed below in detail after which a evaluation will take place.

4.5.1 Word Boundaries

The baseline could not deal with commenter's names that were included in other words, these cases would be identified as if they were referring comments. The straight-forward solution is to tokenize a comment in such a way that it is split up in words and punctuation. Although, than simply looking for a name within the resulting list of tokens would never find names spanning more than one token. However, tokenizing the name in the same way and than matching the list of name tokens to the list of comment tokens would find these names, the method is described in Algorithm 2.

Algorithm 2 Word Boundaries
---
initialize the set of Commenters as $\emptyset$
order Thread ascending by time
for each comment $\in$ Thread **do**
  Tokens $\leftarrow$ tokenize(comment)
  for each commenter $\in$ Commenters **do**
    if commenter_of(comment) $\neq$ commenter **then**
      CommenterTokens $\leftarrow$ tokenize(commenter)
      if CommenterTokens $\subseteq$ Tokens **then**
        assert reacts_on(last_post_of(commenter), comment)
      end if
    end if
  end for
  add commenter_of(comment) to Commenters
end for

The tokenizer used, is part of a dutch POS-tagger, as developed in [4] by E. Tjong Kim Sang. Table 8 gives the results on the same sample as the baseline and the comparison to the baseline.

	recall	δ	precision	δ	F_1	δ
Trouw	-	-	-	-	-	-
AD	1.0000	0.1111	0.9000	0.6419	0.9474	0.5474
Teleg.	0.5000	0.1250	1.0000	0.8636	0.6667	0.4667
Nu.nl	0.5399	0.0675	0.8073	0.2284	0.6471	0.1268
NRC	0.6154	-0.0769	0.6154	0.4389	0.6154	0.3341
	0.6638	**0.0567**	0.8307	**0.5432**	0.7379	**0.3477**

Table 8: Recall, precision and F-measure of the Word Boundaries method, compared to the baseline. The difference, δ, to the baseline is expressed in percentage points. A positive δ indicates improvement, a negative δ indicates deterioration.

Both precision and recall, and therefore the f-measure, show an improvement when the word boundary method is used. The improvement in precision is most striking, this was caused by the fact that the baseline identified many false positives due to falsely identifying references within words. The slight improvement in recall might come as a

surprise, however, this is due the normalization of whitespace conducted by the tokenizer. In the original data whitespaces in the authors name do not always match the whitespaces in the candidates causing the baseline to miss these reacts-on relations.

4.5.2 POS-tagging

A completely different approach would be to start with the comments, identify some candidate words that might refer to other commenters and then try to match each of these candidate to previous authors. This differs from the approach in the baseline method, Section 4.4.2, and the word boundary method, Section 4.5.1, in the sense that it finds candidates that will be discarded afterward and that, given that the method that finds candidates has some sensible heuristic in it, commenters names that resemble commonly used words, will not result in a relation being found. Then, a method to find candidates, that does have some sensible heuristic, needs to be devised. A very sensible heuristic is looking for proper nouns within a comment, these will often indicate a reference to a person or other object. And commenters often use a word as their name that would be recognized as a proper name by POS-tagger. For that reason, a POS-tagger is used to identify candidates within the comments. The used POS-tagger, again, is as described in [4]. The method is shown in Algorithm 3.

Algorithm 3 POS tagging
---
initialize the set of Commenters as $\emptyset$
order Thread ascending by time
for each comment $\in$ Thread **do**
  Tokens $\leftarrow$ pos_tag(comment)
  initialize the set of Candidates as $\emptyset$
  for each token $\in$ Tokens **do**
    if tag_of(token) == Propernoun **then**
      add token to Candidates
    end if
  end for
  for each candidate $\in$ Candidates **do**
    if commenter_of(comment) $\neq$ candidate **then**
      if candidate $\in$ Commenters **then**
        assert reacts_on(last_post_of(candidate), comment)
      end if
    end if
  end for
  add commenter_of(comment) to Commenters
end for

Using this method, we again evaluate the method using the annotated sample set. The result is shown in Table 9.

	recall	δ	precision	δ	F_1	δ
Trouw	-	-	-	-	-	-
AD	0.6667	-0.2222	1.0000	0.7419	0.8000	0.4000
Teleg.	0.3750	0.0000	1.0000	0.8636	0.5455	0.3455
Nu.nl	0.3006	-0.1718	0.8305	0.2516	0.4414	-0.0788
NRC	0.4615	-0.2308	0.8571	0.6807	0.6000	0.3187
	0.4510	**-0.1562**	0.9219	**0.6345**	0.6057	**0.2155**

Table 9: Recall, precision and F-measure of the POS-tagging method, compared to the baseline.

Again there is a great improvement in precision. However, this time the recall drops a little. So whether we should call this method an improvement remains to be seen. The dropping of recall could be the result of some relations that were discovered incidentally by the baseline method.

4.5.3 POS-tagging and loosely-matching

Essentially, the POS-tagging method as described above, boils down to a worse-performing variant of the word-boundary method, Section 4.5.1; names are matched to words in comments but now only those words are considered that have been marked as a candidate, this drops the number of wrong relations, which is desired. However, also the percentage of correct relations that is found drops. Compare table 8 and table 9 for an impression. The power of the POS-tagging method is that it provides candidates; not all words in a comment have to be considered possible references anymore. The change that a candidate should refer to a commenter is much greater than the change that any other word refers to a commenter. This could be exploited by *loosely* matching the authors to the candidates, which could very well improve the number of relations found while maintaining relatively high precision. What is meant by loosely matching needs to be defined, although it will at least involve dropping the string identity requirement. It might therefore accommodate for misspelled names.

The resulting method is very similar to the one shown in Algorithm 3 and is not repeated. The only difference is that instead of checking that $candidate \in Commenters$ holds, the $candidate$ is compared to each $commenter \in Commenters$ using the following similarity measure.

4.5.3.1 Ratcliff/Obershelp similarity.

To match two strings more loosely, the Ratcliff/Obershelp [3] similarity measure is used. This measure is based on the longest common substring that two strings share. It finds the longest substring and then recursively finds the longest substrings on either side of it. When the sum of the lengths of all matching substrings is denoted as $match(w_1, w_2)$ and the sum of the lengths of both strings as $length(w_1) + length(w_2)$ the measure is:

$$similarity(w_1, w_2) = \frac{2 \cdot match(w_1, w_2)}{length(w_1) + length(w_2)}$$

For example, the strings *willem-jan* and *willem* (both dutch names, the second being an abbreviation of the first) have the following similarity measure:

$$similarity(\text{jan-willem}, \text{willem}) = \frac{2 \cdot 6}{10 + 6} = 0.75$$

The measure can be used as a threshold when comparing two strings. A threshold of 0 would match all string to each other. A threshold of 1 would only match identical strings.

4.5.3.2 Result.

Setting the right threshold is not trivial, a low threshold would result in a lot of correctly identified relations but also many relations that should not be found; a low precision. On the other hand, a high threshold would not result in many wrong relations however, not many correct relations would be found either; a low recall. Figure 7 illustrates the correlation between the percentage of correct relations found and the similarity threshold.
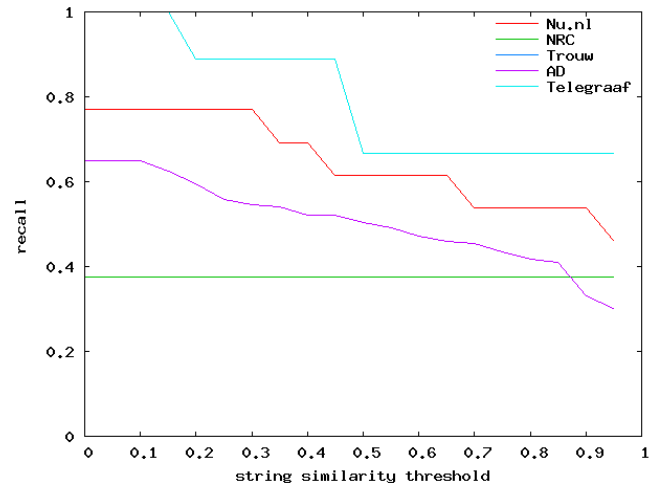

Figure 7: Recall curve for different string similarity thresholds, using the POS-tagging method.
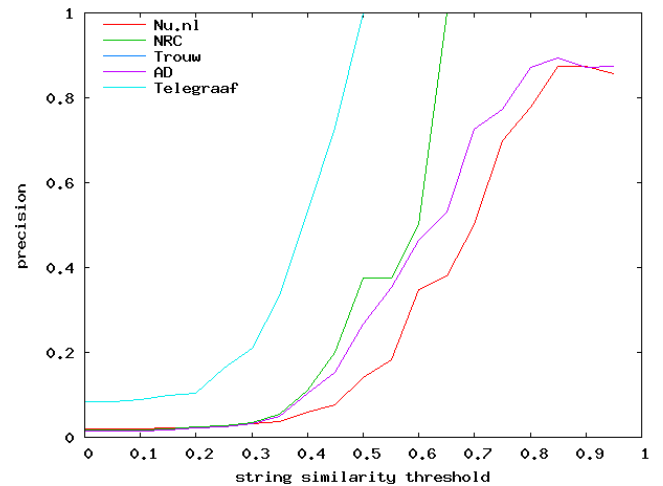

Figure 8: Precision curve for different string similarity thresholds, using the POS-tagging method.

The flip-side is illustrated by figure 8. It can be seen that, although numbers differ per site, the overall trend is the same. It might therefore make sense to take the figure for all sites together, average them, and plot the resulting graphs for both the average recall and average precision into one graph. The averaging is weighted evenly over all sites, so not with a weight based on the number of relations that should be found or the number of comments involved per site, because the resulting method should not have a bias towards the one site with the most comments or relations. The result of this analysis can be seen in figure 9.

If the precision and recall are weighted the same, the intersection would be optimal value to choose. It is not yet decided whet er they are weighted the same, but for now 0.85 is chosen as the optimal threshold since the F_1-measure, that does weight both the same, has an optimum there. The
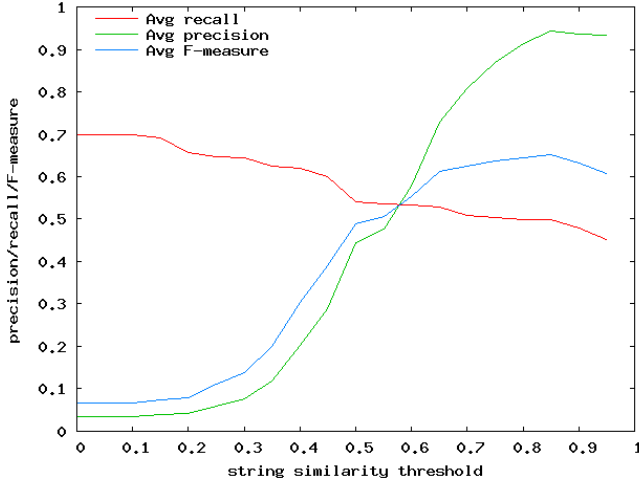
Figure 9: Precision, recall and F-measure curves for different string similarity thresholds, using the POS-tagging method. Here the figures for all sites are averaged.

results are depicted in table 10

	recall	δ	precision	δ	F_1	δ
Trouw	-	-	-	-	-	-
AD	0.6667	-0.2222	1.0000	0.7419	0.8000	0.3999
Teleg.	0.3750	0.0000	1.0000	0.8637	0.5454	0.3454
Nu.nl	0.4110	-0.0613	0.8589	0.2800	0.5560	0.0357
NRC	0.5385	-0.1538	0.8750	0.6985	0.6667	0.3854
	0.4978	**-0.1093**	0.9335	**0.6460**	0.6493	**0.2591**

Table 10: POS-tagging is used to find candidate reference words and these words are loosely matched to commenters with string similarity threshold 0.85, the resulting recall and precision are compared to the baseline

4.5.4 @-Trigger and loosely-matching

Based on evidence found during a manual investigation, Section 4.2, it makes sense to exploit the @-character as a trigger for finding references. The idea is that a word, or words, following a @-character are very likely to refer to a previous author. Therefore, these words are handled as candidates, similar to what was done in Section 4.5.2. After the candidates are identified, they need to be matched to commenters. The resulting algorithm is formalized in Algorithm 4.

As in the POS-tagging case, we can be fairly sure that a candidate does resemble a name of a commenter. In fact, one might think that candidates are in this case even more likely to do this than in the POS-tagging case. This would indicate that a string identity matching between candidate and commenters is not needed, similarity within a certain range, is what is needed. Using the same method as in 4.5.3.2, we come to Figure 10.

As can be seen, the recall is not very high. This is caused by the fact that only a few references are preceded by a @-character, and at some news-sites the @ character is not

---

Algorithm 4 @-Trigger

initialize the set of Commenters as $\emptyset$
order Thread ascending by time
for each comment $\in$ Thread **do**
  Tokens $\leftarrow$ pos_tag(comment)
  initialize the set of Candidates as $\emptyset$
  for each token $\in$ Tokens **do**
    if token $==$ '@' **then**
      for $n = 1..3$ **do**
        add n tokens following @ to Candidates
      end for
    end if
  end for
  for each candidate $\in$ Candidates **do**
    if commenter_of(comment) $\neq$ candidate **then**
      for each c $\in$ Commenters **do**
        if similarity_of(candidate, c) < 0.75 **then**
          assert reacts_on(last_post_of(candidate), comment)
        end if
      end for
    end if
  end for
  add commenter_of(comment) to Commenters
end for
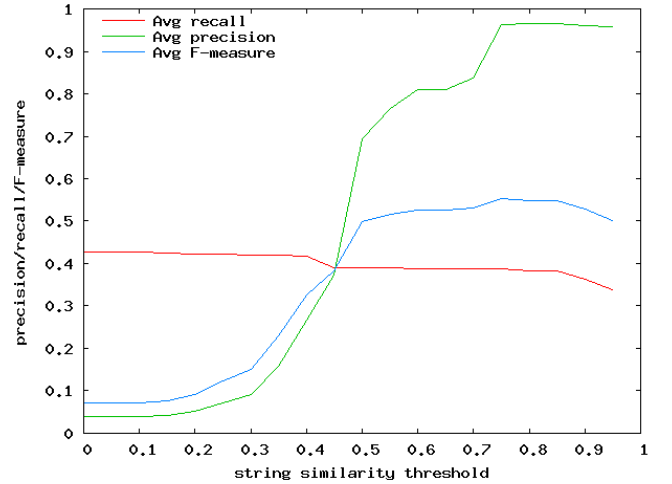
Figure 10: Using the @-character as a trigger for finding candidate references, and then matching these references to commenters, with different string similarity thresholds gives these precision, recall and F-measure curves

used at all, see table 5. However, the slope of the recall curve is not steep, which gives a good ground for opting for a high precision at the cost of just slightly less recall. The optimal string similarity thresholds seems to be 0.75 when the @-character is used as a trigger. Table 11 gives the comparison to the baseline.

4.6 Combining the methods

Now we have three very different methods; the *word boundary* (Section 4.5.1), the *POS-tagging*, with loosely matching

	recall	δ	precision	δ	F_1	δ
Trouw	-	-	-	-	-	-
AD	0.6667	-0.2222	1.0000	0.7419	0.8000	0.4000
Teleg.	0.1250	-0.2500	1.0000	0.8636	0.2222	0.0222
Nu.nl	0.4540	-0.0184	0.7957	0.2168	0.5781	0.0579
NRC	0.3077	-0.3846	1.0000	0.8235	0.4706	0.1893
	0.3883	**-0.2188**	0.9489	**0.6615**	0.5511	**0.1609**

Table 11: Recall, precision and the F-measure for the @-Trigger method with an optimal string similarity threshold of 0.75, compared to the baseline.

(Section 4.5.3) and the *@-trigger* method (Section 4.5.4). The *baseline* (Section 4.4.2) method is regarded as being too simple and error prone to be of any use, other than serving as a ground to base comparison on. And the *POS-tagging*, without loosely matching (Section 4.5.2), is regarded as inferior to the method that does do loosely matching.

The three methods all have a fairly low recall of 0.39 to 0.66, and quite a high precision of 0.83 to 0.95. Knowing that they are all based on different ways of finding candidates and deciding on candidates (see Section 4.6.1.1 on the *word boundary* using candidates), it is not unthinkable that the sets of correctly identified relations they return do not completely overlap. If this were true, combining the three methods would lead to an improvement of the recall while a high precision could be maintained.

4.6.1 Candidates

We have spoken a lot about *candidates* now. At this point it makes sense to give a more precise definition of what a candidate is. All methods that use candidates, consider one comment and search within this comment for words, or a sequences of words, that might refer to a previous author. Essentially, these words, or sequences are what we call *candidates*. Depending on the method used, none, one or even more candidates per comment can be identified. It does not necessarily mean that the relations a candidate indicates in fact *is* a reacts-on relation, it merely means that the method that found the candidate 'believes' that this relation *could* exist. It later decides on whether this relation *should* exist. In both the case of the *POS-tagging* and *@-trigger* methods this deciding is done based on the string similarity threshold.

4.6.1.1 The word boundary method with candidates.

Both the *POS-tagging* and *@-trigger* method are based on the concept of candidates. To make comparison more easily, the *word boundary* method could also be viewed, without change in functionality, as selecting candidates first and then deciding afterward. The candidates here, would be all *n*-grams of words, $n = 1..3$, within a comment, and the decision would be made, based on string identity with a previous author.

4.6.2 Machine Learning

One way of combining the methods, which is similar to the the methodology used in Section 7.5.1 of [1], is to let each of the three methods generate their candidates for the annotated threads, and train a classifier on this data, which then can be used to classify candidates for all threads. The supervised machine learning algorithm, that creates the classifier, needs some features to base its classifier on, these are given below.

4.6.2.1 Feature Set.

- Each **method**, indicated with a number (1 is *word boundary*, 2 is *POS-tagging* and 3 is *@-trigger*), is simply added as a binary attribute. Two or tree methods could come up the same candidate, in which case all method attributes will have the value *true*.

- **Similarity** between candidate and author's name, measured using the Ratcliff/Obershelp similarity, as described in Section 4.5.3.1. Candidates found using method 1 are given a string similarity of 1.

- **Temporal difference** between two posts, measured in the number of comments that separate the comment in which the candidate appears and the comment it might be referring to.

- The **news-site** that is the source of the thread, is added as a nominal attribute indicated with a number (1 is *Trouw*, 2 is *AD*, 3 is *Telegraaf*, 4 is *Nu.nl* and 5 is *NRC*)

- The **word-frequency of the author's name** is looked up in a table containing word frequencies[4]. Of this, we take the $-log$, to account for precision limitations in the software used.

- The **word-frequency of the candidate** is measured equal to the feature above.

4.6.2.2 Creating training data.

Supervised machine learning algorithms take in a table of instances that belong to a certain class. In our case, the class is binary, meaning that an instance can belong to either *true* or *false*, corresponding to respectively, *yes, this should be a relation* and *no, this should not be a relation*. There are, however, some restrictions to what are good sets of instances. There need to be enough instances of all, both in this case, classes, or otherwise the classifier would simply mark all instances with the class of the majority, still yielding a high f-measure. So we need to provide counter examples as well. This could be done by letting go of the found 'optimal' thresholds for the *POS-tagging* and *@-trigger* method and using a wider range instead. Ranging the similarity from 0.5 to 1.0 gives us much more candidates, of which many are counter examples. Also, for the *word boundary* method, more candidates could be found if *n*-grams with $n = 1..4$, or even larger, are used.

Then, how should we attach the right class to the instances in the train set? This is straight forward. If, for a comment, some candidates seem to refer to an author, and if the annotated data confirms that there should be a reacts-on relation, then assign the candidate with the highest similarity to the class *true*. All other candidates get assigned to the class *false*. Using this method, we acquired a set of 986 instances of which 144 belonged to the class *true*. A small sample of the train data is given in Table 13.

---

[4]The frequency list was extracted from a dutch corpus dating back to December 2002, by E. Tjong Kim Sang, enriched with data extracted from all comments considered, to account for changes in language.

Meth.1	Meth.2	Meth.3	Sim.	Dis.	Site	A. Fr.	C. Fr.	Class
TRUE	FALSE	FALSE	1	7	4	13.21	13.21	TRUE
FALSE	TRUE	FALSE	0.57	162	4	12.02	10.7	FALSE
FALSE	TRUE	FALSE	0.55	8	5	15.6	15.6	FALSE
FALSE	TRUE	FALSE	0.67	37	4	13.99	9.4	FALSE
FALSE	TRUE	FALSE	0.57	44	5	13.3	10.81	FALSE
FALSE	TRUE	FALSE	0.86	38	4	13.41	12.31	TRUE
FALSE	FALSE	TRUE	0.46	170	4	13.81	13.21	FALSE
FALSE	FALSE	TRUE	0.46	1	3	15.6	15.6	FALSE
FALSE	FALSE	TRUE	0.55	55	4	15.6	10.92	FALSE
FALSE	TRUE	FALSE	0.57	51	4	11.99	10.31	FALSE
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 12: A snipped of the table that was fed to the machine learning algorithm.

4.6.2.3 Training a classifier.

As a wide range of supervised machine learning algorithms is available, choosing which one to use is a matter of requirements and sensible experimenting. A requirement could be that we want the output to be easily readable and self explaining. Two types of algorithms that have this property are tree-learners and rule-learners. The J4.8 algorithm [5], an implementation of C4.5 [2], with its default settings, gave the best results. Experimenting with less insightful methods, like Bayesian networks and neural network methods, showed that very similar result[5] are obtained, so there seemed to be no reason not to use J4.8 as the classifier.

Since J4.8 is a tree-learner, it produces a decision tree. The resulting tree is depicted in Figure 11. This tree was evaluated on the train set, using 10-fold cross-validation which gave the results in Table 13.

TP Rate	FP Rate	Precision	Recall	F$_1$	Class
0.872	0.021	0.878	0.872	0.875	true
0.979	0.128	0.977	0.979	0.978	false

Table 13: The accuracy per class, for the final decision tree, as it was measured by the 10-fold cross-validation.

It is interesting to see how each method contributes to the final classifier. In Figure 11, the area in which each method is used is given a colored circle. The *word boundary* method, method 1, is active in the blue circle and identifies the majority (106 out of 144 instances) of the reacts-on relations, based mainly on the word frequency of the candidate and the distance the reacts-on relation spans. Second comes the *POS-tagging* method with 20 instances by simply canceling out very low similarities. Last comes the *@-trigger* method that bases classification mostly on candidate and author word frequency.

Clearly, the result of the cross-validation on the train set, Table 13, gives a highly optimistic evaluation because not all relations that should be found were included in the train set, since the methods are not able to identify all desired candidates, some are left out.

What we need, is simply the evaluation method used throughout this paper. So, now using all three the methods, just as they were used to construct the train set in Section 4.6.2.2, create many candidates and their features, Section 4.6.2.1, and classify these using the decision tree. The result of this can be compared to the baseline resulting in Table 14.

---

[5]Only when tweaking the settings of the Multilayer Perceptron marginal improvement was found.

	recall		precision		F$_1$	
		δ		δ		δ
Trouw	-	-	-	-	-	-
AD	1.0000	0.1111	1.0000	0.7419	1.0000	0.6000
Teleg.	0.5000	0.1250	1.0000	0.8636	0.6667	0.4667
Nu.nl	0.6810	0.2086	0.8672	0.2882	0.7629	0.2426
NRC	0.6923	0.0000	0.8182	0.6417	0.7500	0.4687
	0.7183	**0.1112**	0.9489	**0.6615**	0.8177	**0.4275**

Table 14: Precision, recall and F-measure for the combined methods, compared to the baseline.

4.7 Evaluation

It can be easily seen that the three methods indeed all contribute to the final method, in fact, the decision tree that was found by the machine learning algorithm did even use the fact a candidate was found by a certain method as an attribute to base decisions on.

	recall		precision		F$_1$	
		δ		δ		δ
Baseline	0.6071	-	0.2874	-	0.3901	-
Word	0.6638	0.0567	0.8307	0.5432	0.7379	0.3477
POS	0.4510	-0.1562	0.9219	0.6345	0.6057	0.2155
POS loose	0.4978	-0.1093	0.9335	0.6460	0.6493	0.2591
@-Trigger	0.3883	-0.2188	0.9489	0.6615	0.5511	0.1609
Combined	0.7183	0.1112	0.9489	0.6615	0.8177	0.4275

Table 15: Precision, recall and F-measure for all methods, compared to the baseline. These figures are all averaged over all sites (with Trouw left out).

Table 15 shows all methods compared to each other. It can be seen that the highest recall is measured when the methods are combined. This is not a big surprise. It is more of surprise that also the precision is the highest found, together with the @-Trigger method. It is very possible that adding even more methods the combination process would improve the result even more, although there is, evidently, an certain upper-bound in what can be achieved; certain misspellings will lead to ambiguities which require semantics to be solved.

5. FUTURE WORK

As was written in the Introduction of this paper, it might be of interest to social scientist to see if there is a debate going on about a certain subject. As newspaper sites attract not only Internet inclined people, this debate might somehow reflect the debate that is going on in a society. To implement this, some measure of 'liveliness' of discussion would need to be devised and furthermore, articles would need to be grouped by a events. With events we mean a certain real-world event, about which many articles, across news-sites, report. Grouping articles this way, and measuring liveliness in the attached discussion threads would give an indication of how 'hot' a certain event is.

We spoke a few times about different types of reacts-on relations. These types were left out the conducted analysis since it would possibly require looking into the semantics of post. However, it is not unthinkable that the use of certain words could also help identifying the type of relation. Questioning, for example, might involve using many words like *what, how, when*, etc. A post that is an agreement, to give another example, might involve lots of 'positive words' like
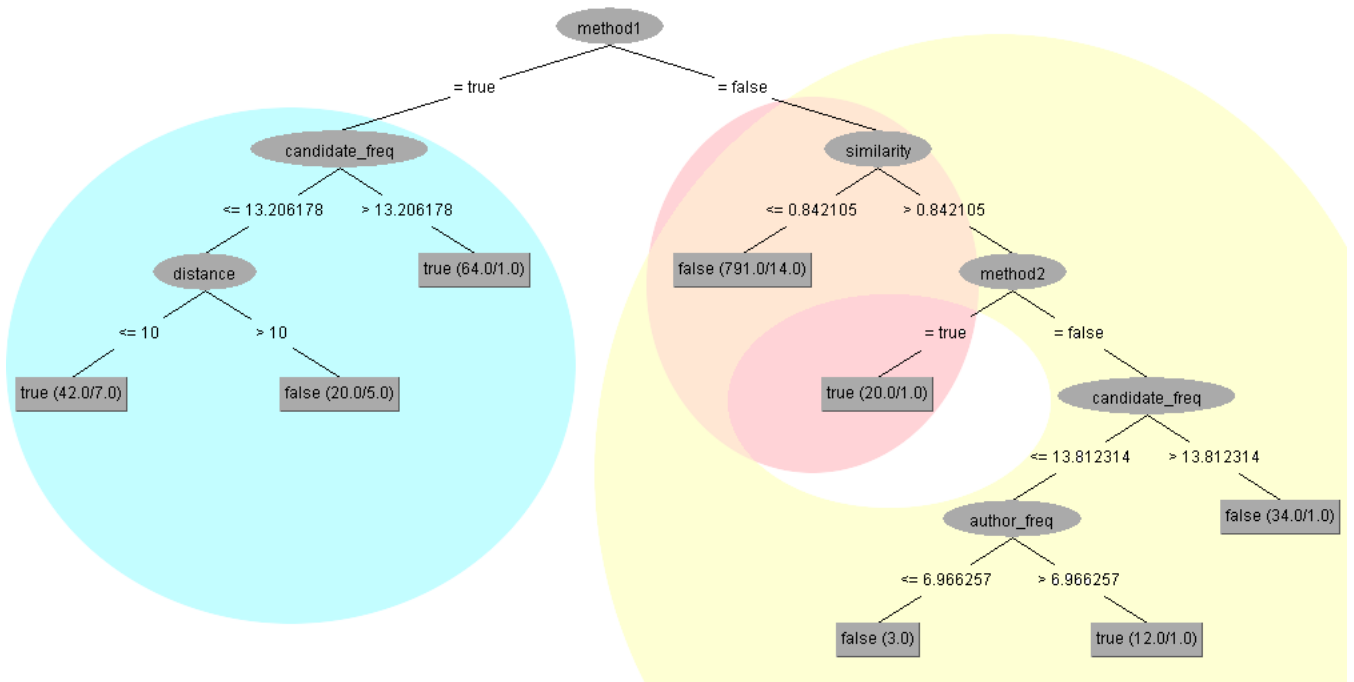
Figure 11: A decision tree, using various features (as described in Section 4.6.2.1), that decides whether a relation between two posts is present or not. Also, the areas where different methods are used are depicted by colored circles. The *word boundary* method is blue, the *POS-tagging* method is red and the *@-trigger* method is yellow.

yes, *indeed*, *agree*, etc. Knowing the type of reaction would be very valuable when trying to enrich the data of articles and could also improve the 'liveliness' measure, mentioned above.

The reacts-on relation, as it is now, could readily be used by the news-sites themselves to improve their interface. It would be useful to add hyperlinks to both posts reacting on a comment and the post the comment is reacting on. A very simple demo application doing just this was implemented by the author to see how this would work. It showed helpful in following discussion and clicking back and forth feels quite natural. Is should however be noted that focusing on a high precision in this sort of applications is crucial. It is hard to sell that some random word get identified as links, it is much easier to explain that some links were missed due to misspellings and the like.

6. CONCLUSIONS

Section 3, on data exploration, on its own already answers parts of the research questions. It shows what comments look like, where they come from and who is posting them. The next section, Section 4, about thread structure does give answer to *how* structure can be found with quite a high recall and a surprisingly precision. This, and the methods used in Section 3, answer the research question aimed at information science: *How can properties of the 'commento-sphere' be brought to light?*. The methods developed shed light into the commentosphere. However, as was already mentioned as future work in Section 5, the precise details of how the structure of threads look is not completely clear

yet; there is no insight in the type of relations between comments that is found by our methods. This leaves part of the research question un-answered, although a direction for future research is given.

7. ACKNOWLEDGMENTS

I am indebted to my supervisors Maarten Marx and Maarten de Rijke for guiding me and offering the idea for this research in the first place. Maarten Marx should also be thanked for his numerous suggestions for improvements. Also Erik Tjong Kim Sang is thanked for providing tools for language analytics.

8. REFERENCES

[1] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, 2007.
[2] J. Quinlan. *C4. 5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
[3] J. Ratcliff and D. Metzener. Pattern matching: The Gestalt approach. *Dr. Dobb's Journal*, page 46, 1988.
[4] E. F. T. K. Sang. Generating subtitles from linguistically annotated text. Atranos report WP4-12, University of Antwerp, 2003.
[5] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

APPENDIX

A. CATEGORIES

See Table 16 for an overview of the different categories used on sites and their correspondence to categorizations used in this paper.

B. THREAD SAMPLE

Following are 5 random samples of threads with 2 comments or more. They are included to illustrate the phenomena observed in the data.

VS waarschuwen China over naleven mensenrechten from Nu.nl

comment 1

author: Wilma H
city:
date: 2007-06-02 21:10:12

De Verenigde Staten hebben vrijdag een scherpe waarschuwing uitgesproken in de richting van China. Volgens Washington kan het Chinese gastheerschap van de Olympische Spelen van volgend jaar in gevaar komen doordat het land de mensenrechten onvoldoende naleeft.

comment 2

author: Frankster
city:
date: 2007-06-02 21:10:12

Moet je horen wie dat zegt. Amerika die een land veroordeelt vanwege het niet naleven van de mensenrechten????? Guatanamo Bay? Electrische stoel? Gaskamer?

comment 3

author: daniella
city:
date: 2007-06-02 21:10:12

Ik snap sowieso niet waarom een land dat zijn burgers dermate wanhopig maakt dat ze hun pasgeboren dochters dumpen welk internationaal evenement ook gaat organiseren. Ze hebben binnen de eigen grenzen nog zat te organiseren, lijkt me.

comment 4

author: Djieff
city:
date: 2007-06-02 21:10:12

De toenaderingspolitiek van de meeste democratische landen heeft wel degelijk effect. China is snel aan het veranderen, ook wat betreft de mensenrechten.Vergeleken met 10 jaar geleden, kan je zelfs spreken van een vrijheid die daarvoor ongekend was. Er is weer een idioot zoals Bush voor nodig om dingen weer op scherp te stellen en zodoende China te provoceren. Gelukkig zijn de chinesen niet gek en lappen ze de uitspraken van een gesjeesde dictator aan hun laars.

comment 5

author: bringer of torture
city:
date: 2007-06-02 21:10:12

Foreign Category	Site	Category
europa	Trouw	buitenland
nederland	Trouw	binnenland
economie	Trouw	economie
dossiers	Trouw	dossier
nieuws	Trouw	algemeen
sport	Trouw	sport
wereld	Trouw	buitenland
diagnose	Algemeen Dagblad	algemeen
nijnatten	Algemeen Dagblad	lokaal
adcupdenhaag	Algemeen Dagblad	lokaal
zuid	Algemeen Dagblad	lokaal
op_weg_naar_peking	Algemeen Dagblad	sport
holleeder	Algemeen Dagblad	dossier
fcutrechtopcuracao	Algemeen Dagblad	sport
lezersrechtbank	Algemeen Dagblad	juridisch
woerden	Algemeen Dagblad	lokaal
groenehartsport	Algemeen Dagblad	sport
loperswereld	Algemeen Dagblad	sport
voetbalcup	Algemeen Dagblad	sport
familie_en_relaties	Algemeen Dagblad	algemeen
economie	Algemeen Dagblad	economie
amersfoort	Algemeen Dagblad	lokaal
hetnieuwewonen	Algemeen Dagblad	algemeen
waterweg	Algemeen Dagblad	algemeen
amsterdam	Algemeen Dagblad	lokaal
bizar	Algemeen Dagblad	entertainment
buitenland	Algemeen Dagblad	buitenland
groenehart	Algemeen Dagblad	lokaal
cultuurenshow	Algemeen Dagblad	entertainment
sport	Algemeen Dagblad	sport
dewereld	Algemeen Dagblad	buitenland
denhaag	Algemeen Dagblad	lokaal
rotterdamvantoen	Algemeen Dagblad	lokaal
natuurprijs	Algemeen Dagblad	algemeen
videolog	Algemeen Dagblad	entertainment
zoetermeer	Algemeen Dagblad	lokaal
clubpaginas	Algemeen Dagblad	sport
drechtsteden	Algemeen Dagblad	lokaal
plekkenvanherinnering	Algemeen Dagblad	algemeen
cultuur	Algemeen Dagblad	entertainment
reizen	Algemeen Dagblad	algemeen
rotterdam	Algemeen Dagblad	lokaal
columns	Algemeen Dagblad	algemeen
multi-media	Algemeen Dagblad	entertainment
westland	Algemeen Dagblad	lokaal
albers	Algemeen Dagblad	algemeen
clubsfcutrecht	Algemeen Dagblad	sport
voorneputten	Algemeen Dagblad	lokaal
oost	Algemeen Dagblad	lokaal
autowereld	Algemeen Dagblad	auto
delft	Algemeen Dagblad	lokaal
stad	Algemeen Dagblad	algemeen
gouda	Algemeen Dagblad	lokaal
hoeksewaard	Algemeen Dagblad	lokaal
binnenland	Algemeen Dagblad	binnenland
airrace	Algemeen Dagblad	entertainment
alphen	Algemeen Dagblad	lokaal
drechtstedensport	Algemeen Dagblad	sport
rivierenland	Algemeen Dagblad	lokaal
badplaats	Algemeen Dagblad	algemeen
oranjeinazie	Algemeen Dagblad	algemeen
uwgeld	Algemeen Dagblad	economie
borst	Algemeen Dagblad	algemeen
spyker	Algemeen Dagblad	auto
filmwereld	Algemeen Dagblad	entertainment
uwmening	Algemeen Dagblad	algemeen
uruzgan	Algemeen Dagblad	buitenland
utrecht	Algemeen Dagblad	lokaal
dossiers	Algemeen Dagblad	dossier
fun	Algemeen Dagblad	entertainment
closeup	Algemeen Dagblad	algemeen
leidenmarathon	Algemeen Dagblad	sport
rotterdamtentop	Algemeen Dagblad	lokaal
uit	Algemeen Dagblad	entertainment
autonieuws	Telegraaf	auto
telesport	Telegraaf	sport
prive	Telegraaf	entertainment
digitaal	Telegraaf	digitaal
woonkrant	Telegraaf	algemeen
binnenland	Telegraaf	binnenland
buitenland	Telegraaf	buitenland
parlement	Telegraaf	politiek
wirefeed	Telegraaf	digitaal
algemeen	Nu.nl	algemeen
wetenschap	Nu.nl	wetenschap
dvd	Nu.nl	entertainment
muziek	Nu.nl	entertainment
plugged	Nu.nl	entertainment
column	Nu.nl	algemeen
economie	Nu.nl	economie
internet	Nu.nl	digitaal
opmerkelijk	Nu.nl	entertainment
beurs	Nu.nl	economie
sport	Nu.nl	sport
achterklap	Nu.nl	entertainment
film	Nu.nl	entertainment
boek	Nu.nl	entertainment
	NRC	

Table 16: Overview of corresponding categories

China wordt machtig net als Duitsland in WO II. Amerika geeft niks om mensenrechten maar alleen om macht.

comment 6

author: ouwerocker
city:
date: 2007-06-02 21:10:12
Bravo Djieff en Frabkster !!!! 100 % mee eens !!!!! Woon er een maand of 4 per jaar, maar loop al te rillen alleen al aan de gedachte dat ik ooit een bezoekje zou brengen aan the USA, wat een land, ellende en grootheidswaanzin met hun "democratie "

comment 7

author: J.J.
city:
date: 2007-06-02 21:10:12
Eens met "bringer of torture" (wel een rare naam trouwens... je bent toch wel gezond?). Beide partijen zijn vooral in macht geinteresseerd. Niet zo gek, maar de vraag is, wat DOE je met die macht. Beiden zijn er heilig van overtuigd dat zij zelf de "goeien" zijn.. en de ander.. krijgt mischien het voordeel van de twijfel, maar meer ook niet. Er verandert veel in China en de mensen hebben het beter dan voorheen.., maar geldt dat voor ALLE mensen? Ik vermoed dat er vooral een boven/middenlaag is die van de economische vooruit-gang profiteert. Gezondheidsvoorzieningen voor de onder-laag gaan hard achteruit (heb ik gehoord op een bbc documen-taire) en in de tussentijd is China met een enorme militaire opbouw bezig - toevallig of niet wordt buurland Nepal door een Maoistische opstand ontregeld, en Tibet is cultureel en etnisch bijna van de kaart geveegd. Tibetanen die het land uitvluchten worden afgeschoten, gevangen gezet, of verdwij-nen (http://www.youtube.com/watch?v=hKcRTfpogAM). In-ternet wordt gecensureerd (met hulp can Microsoft, Yahoo en Google!) en onwelgevallige journalisten lopen nog steeds een reeel risico gearresteerd en monddood gemaakt te wor-den. Religie mag bestaan (lees: wordt getolereerd), maar het bestaat alleen onder stricte controle en regulering va-nuit Beijing. Bovendien is de communistische ideologie, van het bevrijden van alle "verworpenen der aarde" nog niet verlaten. Kortom, ik ben geenszins gerust op de motieven van het Chinese bewind en vind het zeer ongepast dat de wereld China een wereldwijde propagandastunt kado doet, in de vorm van de Olympische spelen. Het enige voordeel daarvan is welicht, dat sporters en olympische comittees van al die landen nu ook in de gelgenheid gesteld worden om met een massale boycot dit propagande offensief te laten terugslaan. Laten we hopen - en er toe aanmoedigen - dat velen die verantwoordelijkheid nemen. Zodat China ook de kans krijgt om zichzelf te corrigeren en niet in de droom bli-jft leven dat het zo wel goed is, allemaal.

comment 8

author: J.J.
city:
date: 2007-06-02 21:10:12
Dit is de correcte link mbt vluchtende Tibetanen: http://www.youtube.com/watch?v=hKcRTfpogAM

comment 9

author: dr oudman
city:
date: 2007-06-02 21:10:12
VS en mensenrechten...?!!

comment 10

author: Ciguli Abi
city:
date: 2007-06-02 21:10:12
Whuhaaahahahahaaa......... Die Amerikanen hebben echt humor. Dat uitgerekend zij China moeten waarschuwen...... Om je dood je lachen (is ook een vorm van foltering).

comment 11

author: ouwerocker
city:
date: 2007-06-02 21:10:12
J.J. , toch maar eens met eigen ogen gaan zien hoe het daar is..... propaganda wordt ook door de westerse wereld bedreven...

comment 12

author: Kimbo
city:
date: 2007-06-02 21:10:12
De pot verwijd de ketel dat hij zwart ziet

comment 13

author: MauriceO
city:
date: 2007-06-02 21:10:12
Hoe zit het met hun eigen mensenrechten?????

comment 14

author: Buu
city:
date: 2007-06-02 21:10:12
Wat een misselijke bedoeling! Kijk naar CUBA waar zij hon-derde mensen hebben opgesloten! Wie zijn zij feitelijk? De moordernaars van deze planeet, het vuil dat voor de bezem gaat? Stop met die onzin, kijkt eerst na Uw zelf alvorens andere landen te betichten! Voor onms is er maar een mo-gelijkheid en dat is de VS links laten liggen, want zij zijn de moordernaars van 2000 tot op heden! Ga verder met haat zaaien, hopelijk zullen jullie dat ook oogsten!

comment 15

author: J.J.
city:
date: 2007-06-02 21:10:12
¿ door ouwerocker 39 minuten geleden: ¿ J.J. , toch maar eens met eigen ogen gaan zien hoe het daar is..... Lijkt me een goed plan :) Ben er van overtuigd dat in China, net als in de VS, zowel goeie als slechte zaken te vinden zijn. Maar het moet beter. Altijd, aan beide kanten. En om dat te kunnen bereiken moeten we - inderdaad - altijd proberen te kijken naar wat er werkelijk gaande is. En ophouden met vloeken, schelden, beschuldigen en demoniseren. Geloof het

of niet, ook in de VS wonen hele aardige mensen ;-) (en ja, ook hele domme)

comment 16

author: ouwerocker
city:
date: 2007-06-02 21:10:12
J.J. daar kan ik me helemaal in vinden. Een wezenlijk verschil vind ik nog steeds, dat over het algemeen de welvarende landen continu met hun vingertje zitten te wijzen, laten ze eerst eens wat doen in hun eigen landje, wij hebben niet meer het excuus dat we landen in de ontwikkelingsfase zijn..... en er is nog meer dan genoeg te doen Wat Tibet betreft, inderdaad negatieve aspecten, maar vergeet de positieve niet. In de periode dat de Dalai Lama aan de macht was, werd het volk op vreselijke wijze uitgebuit, ook daar is genoeg info op internet te vinden. Geloof en staat gaan nu eenmaal nooit samen, voorbeelden te over, zowel in het verre Oosten als in de Westerse landen en Amerika, zowel in het verleden als het heden. Dat er heel aardige mensen in Amerika wonen, daar ben ik van overtuigd, ken er wel een fiks aantal. Wat mij tegen de borst stuit is de opvatting van "democratie "en hoe zij (regering / beleidsmakers) dat als ware kruisridders de hele wereld door de strot willen duwen. Nou, om die vorm van democratie moet je maar verlegen zitten....

comment 17

author: Frankster
city:
date: 2007-06-02 21:10:12
Het chinese volk heeft lang onder het juk van Mao moeten leven. De huidige partijleider Hu Jintao is wel degelijk hervormingsgezind alleen brengt hij het op een heel gematigde manier. Binnenkort krijgt hij ook de macht over het volksleger en dan zal je zien dat het temp omhoog geschroefd wordt.Nu moet hij het niet te snel willen doen want anders grijpt het leger weer in. Deze man speelt het op een bijzonder voorzichtige manier. Te snel hervormingen doorvoeren dat werkt helaas niet. Kijk bijvoobeeld naar Duitsland en de val van de muur. De mensen hebben daar meer dan 30 jaar in een totaal andere wereld gelefd en worden plotsklaps (Kohl moest zijn populariteit vergroten) met een andere wereld geconfronteerd. Het is daar nu een en al ellende,zeer hoge werkloosheid,extreem rechts tiert welig. Ik denk dat de wereld nog voor grote verrassingen komt te staan voor wat betreft deze Hu Jintao. Wil de volgende generatie bij ons nog hoge functies in het bedrijfsleven willen bezetten dan moeten ze maar snel chinees leren en wat minder engels. Ikzelf wil heel graag een paar jaar naar China,Shanghai om de enorme groei daar eens mee te beleven. De VS? Nee dank je,mij te gevaarlijk. @ouwerocker. Ik sta volledig achter jouw mening. Frank

comment 18

author: Tech
city:
date: 2007-06-02 21:10:12
@Kimbo dat was het eerste wat bij mij opkwam :) Land of the slaves, Home of the snakes. Eerst bij je voordeur beginnen Amerika en dan pas naar de rest van de wereld kijken.

comment 19

author: al
city:
date: 2007-06-02 21:10:12
die amerikanen zijn volgens mij helemaal de weg kwijt dan is je zelfkritiek wel heeel ver te zoeken

comment 20

author: numlock
city:
date: 2007-06-02 21:10:12
m je rot te lachen gaat de gristen terrorist Bush nu ook China aanvallen?

comment 21

author: Grimwinkle
city:
date: 2007-06-02 21:10:12
Hahaha, zeggen de VS die marteling legaal willen maken. En totdat dat zover is nemen ze deze "ondervragingstechniek" vast in gebruik. De Geneefse Conventie, da's wc papier toch? Dat is een beetje hetzelfde als de VS die waarschuwen dat andere landen wel eens een atoombom zouden kunnen ontwikkelen, dat mag niet toegestaan worden! Hiroshima da's een soort sushi toch? Als het niet zo in-triest was dan zou ik er hard om lachen.

comment 22

author: Schlebaum
city:
date: 2007-06-02 21:10:12
Je weet wel: de pot en de ketel.

comment 23

author: gebruiker
city:
date: 2007-06-03 21:10:13
Falun Gong leden worden in China op de meest gruwelijke manieren gemarteld, geaborteerd, verkracht en gebruikt voor hun organen. Daar hoor je BNN niet over. Weten zij veel. Nee want je bent jong en je weet niet veel. De Tibetanen worden wreed onderdrukt. Als we dadelijk een VN-wereldregering krijgen dan lopen hier straks Chinese ME'ers rond, want elk land heeft in een wereldgemeenschap immers zijn specialiteit. Men past zijn specialiteiten aan de smaak van het land aan, op de wijze van de Chinese restaurants. Voor Nederland wordt dat afhalen speciaal (bij alle vestigingen van de AIVD), kielhalen speciaal, radbraken Peking Duck, vierendelen met os (voor vier personen).

comment 24

author: annely
city:
date: 2007-06-03 21:10:13
oh, moeten de Olympische Spelen nu weer in de war wor-

den gestuurd door de VS? De sporters en het gewone Chinese volk hebben toch geen schuld aan wat 18 jaar geleden is gebeurd? Over de duizenden doden in Irak en de gesneuvelde Amerikaanse soldaten, reppen ze met geen woord! Zal 't de VS echt lukken om de Olympische Spelen in China tegen te houden?

Amerikaans meisje gevonden na vermissing van een jaar from Nu. nl

comment 1

author: don-pedro
city:
date: 2007-06-07 06:32:54
*sjeeeeeeeeeee
Wat een enorm ziek stel toch weer
Ook de zelfde straf geven als sadam*

comment 2

author: addie
city:
date: 2007-06-07 07:03:54
wat een gekke

comment 3

author: LadyXochi
city:
date: 2007-06-07 07:32:54
Arm meisje... Ik hoop dat ze de verschrikkingen die ze het afgelopen jaar heeft moeten doorstaan te boven komt... Zo zie je maar weer dat het ook bekenden kunnen zijn :(

comment 4

author: daniella
city:
date: 2007-06-07 08:35:54
Ik ben blij voor de ouders dat deze nachtmerrie voorbij is. Misschien begint het voor die arme meid nu pas, maar met tijd en juiste opvang zal ze het hopelijk te boven komen. Flink aanpakken die engerds...! Ik ben blij dat ik geen kennissen heb.

comment 5

author: ef.jonker
city:
date: 2007-06-07 08:42:54
Gelukkig dat ze vrij is. Daders? Zie Don Pedro.......

comment 6

author: Kraaltje
city:
date: 2007-06-07 10:18:54
Zieke toestanden.

comment 7

author: Simplicity78
city:
date: 2007-06-07 10:27:54
Mooi dochtertje gekregen onlangs dus voor mij geen kennissen.

comment 8

author: Harrie_H
city:
date: 2007-06-07 10:41:54
Misschien had ze wel een heel lekker CONNETICUTje.. (sorry voor lompheid van de grap, maar kon het niet laten)

comment 9

author: Grunn1982
city:
date: 2007-06-07 13:21:54
Ze moeten die kennissen geen doodstraf geven! Ze moeten ze individueel opsluiten in een klein hokje van 1 bij 1 meter die pikdonker is en hun net genoeg eten en drinken geven om ze in leven te houden. En laat ze daar hun hele leven maar lekker verrotten.

comment 10

author: P'Chilla
city:
date: 2007-06-08 12:10:37
Wat een zieke mensen weer... En dit: Het meisje zag er volgens de politie bleek uit.... Duhhh.... :-D

Minder tbs'ers op de vlucht from Telegraaf

comment 1

author: John
city: Amsterdam
date: 2007-06-05 15:07:00
Mooie propaganda, na al die moorden die niet nodig waren.

comment 2

author: lind
city: ginneken
date: 2007-06-05 16:12:00
natuurlijk moet de media dit publiceren want veronderstel dat ze de waarheid schrijven dan zal er door deze gasten weinig meerop verlof worden gegaan! wat ik zelf dan dat het niet nodig is om verlof te krijgen want daar hadden ze van te voren aan moeten denken!!

comment 3

author: K
city: breda
date: 2007-06-05 16:15:00
en hebben ze alle drie er een vermoord voordat ze terug gehaald werden?(das toch een leuk gemiddelde!)

comment 4

author: Hubertus ,Scheveningen
city:
date: 2007-06-05 16:16:00
Wat een geweldige vooruitgang.Gefeliciteerd.Deze levende projectielen horen nooit meer op de burgers te worden losgelaten.Wanneer Albayrak meent dat een TBSer gezond verklaard is,laat ze dan deze "persoon" dan 2 weken op proef bij haar thuis wonen.Maar in geen geval de burgers als proefballonnet misbruiken .

comment 5

author: Leen
city: Almere
date: 2007-06-05 18:58:00
Die cijfers zeggen dus niks, als het aantal verloven met 50 procent gedaald is, zijn er dus mr op de vlucht gegaan.

Marktplaats bereikt mijlpaal from Nu. nl

comment 1

author: m00n
city:
date: 2007-05-30 21:05:04
Goh, wat een feest! Kom daar maar eens mee aan bij alle gedupeerden die bij TROS Opgelicht of Radar hebben aangeklopt. Als vandaag je laptop gejat wordt, staat ie morgen op Marktplaats dat in de praktijk niets anders blijkt te zijn dan een digitaal toevluchtsoord voor allerlei gewiekste criminelen. Van honderd miljoen advertenties blijft weinig over als je alle oneerlijke advertenties schrapt, dus het lijkt me dat hier niet echt een juiste voorstelling van zaken wordt gegeven.

comment 2

author: duval75
city:
date: 2007-05-30 21:05:04
daar ben ik het niet mee eens, ik koop en verkoop ook vaak op marktplaats en het merendeel is helemaal niet oneerlijk. De oneerlijke haal je er over het algemeen zo tussenuit, dus begrijp niet dat er geen beter toezicht is.

comment 3

author: Eb
city:
date: 2007-05-30 21:05:04
Binnen no time de babyspullen op m'n zolder kwijt. Ook een aantal zaken aangekocht. Gaat prima. Geen probleem. Meestal overigens afhandeling door afhalen.

comment 4

author: fctwentefan
city:
date: 2007-06-05 01:57:04
Sorry MooN, Ook ik kan me hier niet in vinden!!! Helaas zul je misschien een keer de dupe zijn geweest van een idioot die marktplaats gebruikt om helers te kweken! Toch ben ik

van mening dat zo'n advertentie-site er niets aan kan doen, dat er ook een klein handje vol advertenties op staan, die niet volgens de regels gelden! Ook ik koop en verkoop wel eens iets via de site, en ik ga ook niet elk product na trekken bij de politie of het wel eerlijke handel is, want daar ga ik gewoon van uit!!! In tegenstelling van er enkel van uit te gaan dat er alleen maar gestolen waar etc. via marktplaats zou worden verkocht!!! Zulke practijken vind ik dat je zo'n site niet mag aanrekenen, daarvoor dien je toch echt bij de adverteerders zelf te zijn!!!

Belgische onderbroek beledigt hindoes from Nu. nl

comment 1

author: CryWolf
city:
date: 2007-06-06 10:50:30
De Brusselse modeketen DOD heeft herenonderbroeken met daarop een afbeelding van de hindoegodin Lakshmi schielijk uit zijn winkels gehaald na protesten van hindoes...

comment 2

author: CryWolf
city:
date: 2007-06-06 10:51:30
Oei..moet ik m'n allah-broeken en Paus-slips ook weer terugbrengen....

comment 3

author: daniella
city:
date: 2007-06-06 11:42:30
O-ver-dre-ven!!! In alle (nou ja, veel) Nederlandse onderbroeken zit een kruis, daar gilt toch ook de hele christelijke gemeenschap niet over? Sterker nog, die gaan pas gillen bij de modelletjes waar ze discreet uit zijn gehaald!

comment 4

author: Ohja
city:
date: 2007-06-06 11:51:30
Onderbroekenlol wordt niet door iedereen gewaardeerd dus.

comment 5

author: nita hardoel
city:
date: 2007-06-06 12:50:30
Vind het belachelijk! geen respect voor een ander geloof! ze moesten maar eens weten wat het woord Lakshmi betekent. zou erg fijn zijn als ze een foto van hun mamske of dochter op zouden plaatsen. Mooi H. stelletje idioten. ze mogen van mij goed gestraft worden.

comment 6

author: isabelle
city:

date: 2007-06-06 23:31:56

ik vind het een beetje dom dat die onderbroek verboden wordt. er zijn toch ook zat t-shirts en onderbroeken waarop Jezus of God belachelijk wordt gemaakt, die zijn toch ook niet verboden?

comment 7

author: Seeker4Life
city:
date: 2007-06-07 14:07:56

Het gaat er niet om of het "verboden "is of niet. Het gaat erom dat dat de fabrikant gewoon niet de moeite had genomen om na te gaan wat het Laxmi plaatje betekent. En ook niet of het geen mensen zou kunnen kwetsen , en dus een rel kon veroorzaken, wat dus ook gebeurde. Zodra hij doorkreeg dat dat wel het geval was, heeft hij de onderbroeken uit de handel genomen, om mensen niet meer te kwetsen. Ja, dit had hij eerder kunnen uitzoeken, maar nee, het is niet "verboden" om dergelijke plaatjes op kledingsstukken te zetten. En natuurlijk is het dom van fabrikanten om teksten en plaatjes op kledingsstukken te drukken waarvan je de betekenis niet weet. Hoeveel mensen kopen wel niet spullen met Chinese karakters of met teksten in het Hindi, puur omdat het "zo exotisch"staat. En dan gewoon niet weten WAT er staat, of wat het plaatje BETEKENT. Alleen in nederland en Belgie is men zo dom, in elk ander land vraagt men de verkoper wat het nu precies betekent.