A comparison of feature selection methods for an evolving RSS feed corpus

Rudy Prabowo,
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street,
Wolverhampton, WV1 1SB, UK. Email: rudy.prabowo@wlv.ac.uk
Tel: +44 1902 518584 Fax: +44 1902 321478
Mike Thelwall,
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street,
Wolverhampton, WV1 1SB, UK. Email: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478

Previous researchers have attempted to detect significant topics in news stories and blogs through the use of word frequency-based methods applied to RSS feeds. In this paper, the three statistical feature selection methods: χ^2, Mutual Information (*MI*) and Information Gain (*I*) are proposed as alternative approaches for ranking term significance in an evolving RSS feed corpus. The extent to which the three methods agree with each other on determining the degree of the significance of a term on a certain date is investigated as well as the assumption that larger values tend to indicate more significant terms. An experimental evaluation was carried out with 39 different levels of data reduction to evaluate the three methods for differing degrees of significance. The three methods showed a significant degree of disagreement for a number of terms assigned an extremely large value. Hence, the assumption that the larger a value, the higher the degree of the significance of a term should be treated cautiously. Moreover, *MI* and *I* show significant disagreement. This suggests that *MI* is different in the way it ranks significant terms, as *MI* does not take the absence of a term into account, although *I* does. *I*, however, has a higher degree of term reduction than *MI* and χ^2. This can result in loosing some significant terms. In summary, χ^2 seems to be the best method to determine term significance for RSS feeds, as χ^2 identifies both types of significant behavior. The χ^2 method, however, is far from perfect as an extremely high value can be assigned to relatively insignificant terms.

Keywords: feature selection, chi-square, mutual information, information gain

1. Introduction

Rich Site Syndication (RSS) is an XML format for publishing concise information updates. It is mainly used by news sites to publish summaries of their latest stories and by Blogs for summaries of their latest postings. Both researchers and commercial companies have noticed that Blogs and RSS feeds have the potential to be used for public-opinion gathering or marketing purposes, and hence there has been a drive to develop effective tools and techniques for the automatic analysis of RSS feeds (e.g., Gill, 2004; Pikas, 2005). Previous researchers have used word/noun/noun phrase time series analysis methods with simple frequency statistics to identify significant topics (Glance et al., 2004; Gruhl et al. 2004). However, there is an established body of research for identifying the most significant words in collections of documents, a task called 'feature selection'. It is therefore logical to assess whether any of the established statistical feature selection methods can help the identification of significant topics in RSS feeds. A first stage of this process is to compare these methods to assess which are the most suitable for this new data source.

In this paper, a 'term' is a noun or noun phrase and a 'feature' is a term that is judged to be significant within a collection of documents. Feature selection methods such as χ^2, Mutual Information (*MI*) and Information Gain (*I*), have been commonly used in different application domains. One example is automatic text classification (Yang & Pedersen 1997): determining document categories based upon a set of significant terms representing the document features (Sebastiani, 2002). The role of feature selection in this context is in condensing documents by removing redundant words, in order to speed document classification without reducing classification quality. Feature selection that is too 'aggressive', in terms of removing too many words, will result in poor document classifications. A very different example is Topic Detection and Tracking (TDT) (Allan et al., 1998b; Yang et al., 1998), which focuses on identifying a new event/topic, and tracking the previously identified event/topic with regard to new incoming stories. This may be achieved by identifying and clustering collections of related terms, but other methods are also used, such as Information Extraction (e.g., Luo, 2004). In comparison to document classification, TDT implicitly requires much more aggressive feature selection because its purpose is to identify significant events across documents rather than to capture the essence of individual documents. A relevant application of TDT is the automatic generation of overview timelines (Swan & Allan 2000) through determining which terms are significant over a given time period.

Despite previous research into term selection methods, there is no clear indication of the superiority of any particular method for all types of data: each has its own strengths and weaknesses. Yang & Pedersen (1997), supported by Sebastiani's (2002) automatic text categorisation review article, suggest that *I* and χ^2 have similar performance in supporting document classification, with both being significantly better than *MI*. Of these, χ^2 seems to be gaining support in practice for classification, perhaps for its calculation efficiency (e.g., Ng et al. 1997) although selecting terms by using χ^2 has also been criticised for being unreliable when the cell value is less than 5 (Dunning, 1993). There is no strong evidence, however, to suggest which method is the most effective for the less studied TDT task of selecting significant terms from document collections, although Swan and Allan (2000) have adopted χ^2 for this purpose. It is not clear that methods which work best for document classification also work best for TDT, for example because for classification purposes it can be useful to eliminate terms based upon a high degree of association with remaining terms, which is not a consideration for TDT (Swan & Allan 2000). Moreover, each method uses a probabilistic function based upon assumptions about the distribution of the data, such as independence, which are violated in practice to varying degrees (Cooper, 1995). Hence experiments are required with each new type of data source and for each new type of task to assess the strengths and weaknesses of the leading methods in practice.

In this paper the Mutual Information (*MI*), χ^2 and Information Gain (*I*) feature selection methods are evaluated for an evolving RSS feed corpus in order to decide which is the most suitable for identifying features that are significant across a number of documents within the collection (the TDT type of task). In particular, the suitability of the three methods for selecting significant features on a given date is assessed. Term Strength (TS) and Document Frequency (DF) Thresholding are significantly different from the other three methods, as these two methods only consider the document space, rather than individual parameters, such as category or date, and require a training corpus. TS, in particular, requires computationally expensive document clustering (Yang & Pedersen 1997). For these reasons, they were rejected as inappropriate for this evolving data set.

In addition, the assumption that very large values indicate highly significant terms is investigated; and the extent to which the three methods agree with each other about the significance of a term on a given date. This is particularly relevant for TDT. An evaluation method that is inspired by the way in which conference papers are reviewed is also proposed and used. A paper may be accepted for publication if two or three referees put a good recommendation. For the evaluation of the feature selection methods used, the extent to which the three methods agree with each other about the degree of the significance of a term on a given date is determined.

This paper is organised as follows. Section 2 reviews the existing work in the area of automatic text classification, topic detection and tracking, automatic generation of overview timelines and blog text analysis. Section 3 describes the type of data used for the evaluation, and the way the data was collected and processed. Section 4 describes the feature selection method evaluation procedure and presents the evaluation results. Section 5 presents the conclusions.

2. Related Work

This section describes research relevant to feature selection in the areas of automatic text classification, Topic Detection and Tracking and overview timeline generation. It also reviews current blog text analysis.

2.1 Automatic Text Classification

A number of researchers have proposed various techniques to carry out an automatic classification task, including: Hierarchic Classification (Jenkins et al., 1999; Larkey, 1998); Neural Network Based Classification (Ng et al., 1997; Chen et al., 1994; Yang, 1994); K-Nearest Neighbour Classification (Yang, 1999; Kwon & Lee, 2000); and Support Vector Machine Based Classification (Dumais & Chen, 2000). There are two main issues for automatic classification.

1. Selecting appropriate features for each category and document. Some researchers use only one particular feature selection method and intuitively set a threshold to exclude terms.
2. Assigning documents to appropriate categories. This task can be hard for a large number of categories and a deep hierarchical classification scheme.

The evaluation strategy that is commonly used to evaluate feature selection methods is to use an application as a means to measure the effectiveness of each method. Yang & Pedersen (1997) have used an automatic text classifier to evaluate the effectiveness of three feature selection methods. For researchers who do not have an automatic text classifier and a collection of pre-classified samples, this strategy incurs an implementation cost and requires human resources to pre-categorise a collection of documents. In contrast, Swan & Allan (2000) use a clustering algorithm to evaluate the effectiveness of the χ^2 method. This strategy also incurs an implementation cost and requires human resources to evaluate the quality of the clusters that are automatically generated by the clustering algorithm. In addition, the parameters and algorithm used in the application can influence the evaluation result. The evaluation result reflects not only the effectiveness of the feature selection methods, but also the performance of the application used.

2.2 Topic Detection and Tracking

In the context of Topic Detection and Tracking, an event is defined as something that happens at a particular time and place (Allan et al., 1998b; Yang et al., 1998); a topic is defined as a seminal event or activity, along with all directly related events and activities (NIST Speech Group, 2005). The term 'story' is often used to describe the natural units of text in which the information arrives, such as a single newswire report. In this context, a set of events can be grouped into a number of topics. Each topic becomes the abstraction (or generalisation) of its associated events. Given a collection of stories, either in form of texts or speech transcripts, TDT encompasses the following issues.

1. Story segmentation: Segmenting a stream of text into its constituent topically cohesive stories (Allan et al., 1998a).
2. Event detection: Identifying new stories that discuss an event that has not been reported in previous stories (Allan et al., 1998b).
3. Event tracking: Keeping track of new stories that discuss a previously identified event (Allan et al., 1998b).
4. Topic detection: Clustering stories which are topically related (Walls et al., 1999);
5. Topic tracking. Keeping track of new stories that discuss a previously identified topic (Jin et al., 1999).

Event detection task consists of two parts.

1. Given a set of stories:
   (a) Assign each term (noun or noun phrase) a weight.
   (b) Select the most significant terms within each story.
   (c) Store the weighted terms as a vector that is regarded as the representation of the content of a story.
2. Given a set of term vectors, stories that discuss the same event are clustered together. The output is a number of clusters. Each cluster contains a set of term vectors that represent an event (Yang et al., 1998; Allan et al., 1998b).

The event tracking task is only concerned with incoming new stories, and consists of three parts.

1. Given a new story,
   (a) Assign each term a weight;
   (b) Select the most significant terms within the new story;
   (c) Store the weighted terms as a vector which is regarded as the representation of the content of the new story;
2. Compute the similarity between the term vector representation of the new story and all the term vectors of the existing stories;
3. Assign the most appropriate event that the new story discusses, if any. Otherwise, a new event is detected.

The topic detection and tracking tasks (Schultz & Liberman, 1999; Walls et al., 1999; Jin et al., 1999) are quite similar to the event detection and tracking tasks. The main difference is that the former focuses on the identification of topics across stories, whilst the latter focuses on the identification of events.

2.3 The Generation of Overview Timelines

Given a set of stories, Swan & Allan (1999, 2000) and Swan and Jensen (2000) have built a $\chi 2$-based model to select features for a given time period, clustering together closely related features. Each cluster represents a topic. This is less complex than the TDT approach, described in section 2.2. Table 1 illustrates the difference in sequential process between the two approaches, in respect to a topic detection task.

Given a collection of stories, both approaches can be used to identify a number significant terms which can be used to identify a number of topics. As shown in Table 1, there is a fundamental difference between the two approaches with overview timeline generation being closer to current blog analysis in terms of feature selection objectives. Topic detection focuses on the generation of clusters, each of which contains a set of stories representing a topic. In contrast, the generation of overview timelines focuses on the selection of significant terms on a certain time period. The degree of the significance of a term is determined with respect to a certain time period. Table 2 and 3 highlight the strengths and weaknesses of each approach.

2.4 Blog Text Analysis

Existing blog text analyses focuses on the extraction of useful information from a set of blogs. The texts found within the blogs are analysed for the following purposes.

- Determining the most significant keywords and proper names within a certain time period. Glance et al. (2004) focus on developing automatic trend discovery across blogs. Their system can detect key persons, phrases and paragraphs on a daily basis.
- Predicting the propagation of a topic. Gruhl et al. (2004) simulate a random graph (Erdös & Rényi, 1960) to predict the spreading of a topic across blogs.
- Organising a set of terms so that concept-based topic identification can be carried out. Given a collection of texts, Avesani et al. (2005) set out to develop a system which can generate a topic-centric view so that bloggers can find related blog entries with respect to a specific topic.

The above initiatives all exploit textual data to find and track some useful information, such as a topic and the name of a person/an organization. They are closely related to TDT, but also different in important respects, for example allowing analyses of the spreading of a topic between bloggers. Current feature selection methods tend to be based upon simple term frequency time series and, like overview timeline generation, focus on identifying individual significant terms (i.e. aggressive feature selection).

3. Data Collection and Processing

In recent years, bloggers, the people who create and edit weblogs (blogs), have used weblogs as a means for publication and a communication medium (Glance et al. 2004). A blog may have a corresponding RSS feed which contains all the blog postings in RSS format. RSS is an XML-based metadata syndication format (Hammersley, 2005). In this context, the term, syndication, refers to the use of an RSS feed as a medium for publishing and sharing information. Despite the fact that RSS feeds have grown more slowly than blogs, perhaps due to a number of RSS specification changes, RSS technology has been adopted by a number of newspaper agencies as a syndication technology at a brisk pace (Gill, 2005). In RSS terminology, an RSS element only contains one sub-element, a 'channel'. A channel element contains information about the RSS channel and a set of items. Each item refers to a single blog posting or other coherent unit of information. An RSS item is used here as the basic unit of analysis, analogous to the TDT 'story'. The RSS analysis experimental procedure is described below.

Mozhdeh: RSS Feed Collector and Data Monitoring System. Mozhdeh (Thelwall, Prabowo, & Fairclough, 2006) was used for collecting data, and has been implemented and maintained by the second author. As a starting point, the system automatically collected 19 587 RSS feeds from Google, and blogstreet (http://www.blogstreet.com/). The system monitored this set of feeds hourly (daily for infrequently updated feeds) and stored each new incoming RSS item found. For each new RSS item, the system recorded the parent feed.

Note that since the findings from this experiment were based upon a single evolving RSS feed corpus they are thus indicative rather than definitive. As with many internet technologies (e.g. Egghe, 2000), RSS is evolving and the uses to which it is put are also evolving beyond news feeds and blogs. Hence it is not possible to find definitive time-independent optimal methods for any type of RSS analysis, nor is it appropriate to set up a static, classified RSS test corpus against which to benchmark the different methods.

Text Extractor. For each item, the Text Extractor extracted the title, description and publication date and then converted the publication date into its associated GMT time. A date converter program was implemented that can converted different types of timestamp formats and different timezones into GMT times, by analysing the timestamp patterns. The total number of timezones used was 63 (http://www.lns.cornell.edu/public/COMP/krb5/krb5-admin/kadmin-Time-Zones.html, accessed at 23 December 2004). The total number of timestamp patterns used was 36. The title and description texts were then stored in a text file.

POS Tagger (Brill, 1994) and **NP Chunker** (Ramshaw & Marcus, 1995). For each text file, a POS tagger and NP chunker were used to tag and chunk the texts to determine the nouns, noun phrases and proper nouns found in the text files.

Inverted File Builder. Given a set of chunked text files, the Inverted File Builder built three indexes, storing each one as a two column database table.

1. TermItem. Used to determine to which item a term belongs;

2. ItemRSSFeed. Used to determine to which RSS feed an item belongs;
3. ItemDate. Used to determine when an item was posted in GMT time (i.e., publication date).

Feature Selector. The Feature Selector was implemented with each of χ^2, Mutual Information and Information Gain. The Feature Selector assigned a value to each term from the RSS feed data on each day during which the term occurred.

4. The Evaluation of the Three Feature Selection Methods

This section describes how χ^2, Mutual Information (*MI*) and Information Gain (*I*) values were computed and used for feature selection (section 4.1) and how the three methods were evaluated as ranking functions, i.e. the extent to which the three methods agreed with each other about term significance on a given date (section 4.2).

4.1 Computing χ^2, *MI* and I values

4.1.1 2x2 Contingency Tables

Prior to calculating χ^2 and *MI,* a 2x2 contingency table was required. Given a term, $term_i$ and a publication date, $date_j$, a 2x2 contingency table was constructed, where:
a is the number of items which contain $term_i$ and are posted on $date_j$;
b is the number of items which do not contain $term_i$ and are posted on $date_j$;
c is the number of items which contain $term_i$ and are not posted on $date_j$;
d is the number of items which do not contain $term_i$ and are not posted on $date_j$.
For each term, a 2x2 contingency table is constructed to determine the observed frequencies, *O*. A second set of 2x2 contingency tables was also constructed for the expected frequencies, *E*.

4.1.2 Computing χ^2 values

Given two sets of 2x2 contingency tables, described in section 4.1.1, the χ^2 value of a term, $term_i$ with respect to a publication date, $date_j$ was computed as follows:

$$\chi^2(term_i, date_j) = \sum_{k=a}^{d} \frac{(O_k - E_k)^2}{E_k} \tag{1}$$

k = {a . . . d} represents the value of each cell in a 2x2 contingency table. A Yates continuity correction was applied to each χ^2 calculation as the degree of freedom is 1. The χ^2 calculation used in this experiment does not approximate the χ^2 value, such as described in Yang & Pedersen (1997) and Swan & Allan (2000). The larger a χ^2 value, the stronger the evidence to reject the null hypothesis, which means that $term_i$ and $date_j$ are dependent on each other, i.e. the proportion of items containing $term_i$ on $date_j$ was significantly higher or lower than elsewhere in the corpus.

4.1.3 Computing Mutual Information (*MI*) values

Given a set of 2x2 contingency tables containing the observed frequencies, *O*, described in section 4.1.1, the *MI* value of a term, $term_i$ with respect to a publication date $date_j$ was computed as follows:

$$MI(term_i, date_j) = log_2 \frac{P(term_i \wedge date_j)}{P(term_i) \cdot P(date_j)} = log_2 \frac{a \cdot N}{(a+b) \cdot (a+c)} \tag{2}$$

Here N is the total number of items used in the collection (= a + b + c + d). The larger an *MI* value, the greater the association strength between $term_i$ and $date_j$, where *MI* ($term_i$, *date*j) must be > 0. This means that the joint probability, P($term_i$, $date_j$) must be greater than the product of the probability of P($term_i$) and P($date_j$). Two examples of the use of this method for measuring the strength of two terms association can be found in Conrad & Utt (1994) and Church & Hanks (1989).

4.1.4 Computing Information Gain (*I*) values

The information gain, *I* (*D*;*T*) is computed based on an entropy value of the date *D* and the conditional entropy of the date *D* given the term, *T*.

$D = \{d, \bar{d}\}$ is a date variable which represents two events: the presence and absence of the Date, *D*.

$T = \{t, \bar{t}\}$ is a term variable which represents two events: the presence and absence of the Term, *T*.

I (*D*;*T*) is then defined as follows:

$$I(D;T) = H(D) - H(D \mid T)$$

$$I(D;T) = \left[-\sum_{j=d,\bar{d}} p(j) \cdot log_2 p(j) \right] - \left[p(t) \cdot \sum_{j=d,\bar{d}} H(j \mid t) + p(\bar{t}) \cdot \sum_{j=d,\bar{d}} H(j \mid \bar{t}) \right]$$

$$I(D;T) = -\left[p(d) \cdot log_2 p(d) + p(\bar{d}) \cdot log_2 p(\bar{d}) \right] - \left[p(t) \cdot \sum_{j=d,\bar{d}} H(j \mid t) + p(\bar{t}) \cdot \sum_{j=d,\bar{d}} H(j \mid \bar{t}) \right] \tag{3}$$

The conditional entropy of *D* given *T* is computed as follows:

$$\sum_{j=d,\bar{d}} H(j \mid t) = -\left[p(d \mid t) \cdot log_2 p(d \mid t) + p(\bar{d} \mid t) \cdot log_2 p(\bar{d} \mid t) \right] \tag{4}$$

$$\sum_{j=d,\bar{d}} H(j \mid \bar{t}) = -\left[p(d \mid \bar{t}) \cdot log_2 p(d \mid \bar{t}) + p(\bar{d} \mid \bar{t}) \cdot log_2 p(\bar{d} \mid \bar{t}) \right] \tag{5}$$

In information theory (Shannon & Weaver, 1963), *I*(*D*; *T*) is used to measure the average reduction in uncertainty about *D* that results from learning the value of *T*. This is the average amount of information that *T* conveys about *D* (MacKay, 2003). *H*(*D*|*T*) represents the amount of reduction in uncertainty. The smaller is *H*(*D*|*T*), the greater the information that can be learned from T. For *H*(*D*|*T*) = 0, the degree of uncertainty in learning the value of *D* is 0. This means that both *I*(*D*; *T*) and *H*(*D*) represent the same amount of information about *D*.

In this paper, the notion of uncertainty was applied to determine the degree of closeness between *D* and *T* by learning the presence and absence of *T* in each RSS item. In this context, the *H*(*D*) of one particular publication date was important rather than the entropy, *H*(*D*) of all publication dates. The presence and absence of the publication date was taken into account, and *H*(*D*|*T*) was used to quantify the degree of uncertainty that the term, *T* was significant on the date *D*. The larger the *I*(*D*; *T*) value, the higher the degree of certainty that *T* was a good indicator for *D*.

Three examples of the use of this method can be found in Yang & Pedersen (1997), MacKay (2003) and Moore (2003). Confusingly, Information Gain, *I* is sometimes called Mutual Information. In this paper, Mutual Information, (*MI*), however, refers to the degree of the association strength between a term and a date without the use of entropy, as explained in section 4.1.3.

4.2 Evaluating χ^2, *MI* and I

As discussed in the introduction, the evaluation method used was inspired by the way a journal paper is reviewed. In this context, the extent to which the three methods agreed with each other on determining the degree of the significance of a term on a given date was investigated.

4.2.1 Evaluation Method and Procedure

The evaluation procedure comprised four steps.

1. Three sets of significant terms were selected, one set for each feature selection method.
2. Two criterion sets of significant terms were generated from the above three sets. The first set, the *3-votes set*, contains the terms that all three feature selection methods agree are significant. The second set, the *2-3-votes set*,

contains the terms that at least two feature selection methods agree are significant.

3. Each set of significant terms (step 1), was compared with the two criterion sets (step 2). The two criterion sets are used as the criteria against which the performance of each feature selection method could be assessed.
4. The performance of each method was then measured in terms of
   (a) *agreement*. The feature selection method and a criterion set agree with each other that a term is significant.
   (b) *disagreement*. The method selects a term as significant, but a criterion set does not include it.
   (c) *miss rate*. The criterion set includes a term, but the method does not select it.

4.2.2 Limitations

The evaluation method does not employ human judgement to determine the effectiveness of the three methods as ranking functions. This leads to the following limitations.

1. The fact that the methods and a criterion set agree with each other does not necessarily mean that the methods have correctly selected a significant term. This is especially significant for the 2-3 votes set. Thus, the term 'agreement' is used instead of 'precision'.
2. The fact that a criterion set disagrees with a method does not necessarily mean that the method incorrectly selected a significant term. This is especially significant for the 3 votes set. Thus, the term 'disagreement' is used instead of 'error'.
3. The term, 'miss rate', can only be interpreted from the criterion set perspective. From a human judgement point of view, it does not necessarily mean that the methods miss genuinely significant terms.

4.2.3 Evaluation Results

The total number of unique terms identified was 1 736 715, and the total number of term-date pairs was 2 912 581, each having a χ^2, *MI* and *I* value. Each method had different lower (LV) and upper values (UV); χ^2: 1.89E-10 – 715 895.8, *MI*: -7.85 - 19.65*, I*: 7.9E-17 - 0.01. To ensure a fair evaluation, the term-date pairs were sorted by their χ^2, *MI* and *I* values, and the same proportion of term-date pairs were selected for each feature selection method. The evaluation procedure discussed in section 4.2.1 was then applied for 39 different levels of reduction (from 61.56% until 99% reduction). By doing this, the proportion of agreements, disagreements and the miss rate for different degrees of significance could be evaluated. Table 4 lists the 39 reduction levels. Here, "x % reduction" means that only the highest (100-x)% term-date pairs were selected. For example, Table 4 (1st row) means for 61.56% reduction, the percentage of term-date pairs selected was 38.44% (= 38.44/100 * 2 912 581 = 1 119 596), and the associated threshold values were 6.64 for χ^2, 3.93 for *MI* and 4.83E-6 for *I*.

Figure 1 shows the evaluation results with respect to the 3-votes set. Figures 2-4 show the evaluation results with respect to the 2-3-votes set. In Figure 1, no miss rate is depicted, since the miss rate = 0. Figure 1 shows that the proportion of agreements between the three feature selection methods and the 3-votes set gradually decreases after a 71% reduction, and sharply decreases after a 97% reduction. At 89% reduction, the proportion of agreements (48%) is less than the proportion of disagreements (51.55%). This experimental evidence suggests that few terms which are assigned extremely high values may actually be less significant than some other terms assigned lower values, as the three methods show a strong disagreement over judging term significance. From figures 2-4:

- χ^2 has the largest proportion of agreements, and the smallest proportion of disagreement with the 2-3 votes set.
- The proportion of agreements sharply decreases after a certain level of reduction (96% for χ^2 and *MI*, and 89% for *I*);
- The proportion of disagreements and miss rates overlap each other frequently;
- χ^2 has an average miss rate of 0.32% with respect to the 39 reduction levels. *I* 16.43%, and *MI* 20.40%.

Despite the use of 2-3 votes sets, which are less stringent than 3-votes sets, the evidence also suggests that the three methods show strong disagreement for a high level of reduction. In addition, the miss rates indicate that *MI* and *I* are less effective than χ^2 for the aggressive exclusion of terms, as *MI* and *I* seem to be more aggressive than χ^2. This is due to the following reasons. The 2-3-votes sets are dominated by χ^2, which is supported by either *MI* or *I*. Hence, χ^2 can have the largest proportion of agreements and the smallest proportion of disagreements and miss rates.

In addition, figure 5 shows the proportion of agreement between two methods only, i.e. between χ^2 and *I*, χ^2 and *MI, MI* and *I*, which means that the 2-3 votes and 3-votes set were not used for comparison. Hence, we can see which pairs agree/disagree. Figure 5 shows that χ^2 and *I* agree with each other the most, although χ^2 and *MI* also show some levels of agreement. In contrast, *MI* and *I* show a strong sense of disagreement from 72% reduction onwards. Moreover, despite some small fluctuations, the proportion of agreements for both pairs χ^2 - *I* and *MI* - *I* gradually decrease. The proportion of agreements between χ^2 and *MI*, however, shows a fluctuation from 65.01% to 83.03% for the reduction levels between 91% and 96%. This fluctuation indicates that for a high level of reduction, χ^2 and *MI* tend to agree with each other.

In addition, the extent to which the size of the corpus used affects the findings, as illustrated in figures 1–5, is important. The first 10 000 RSS items were selected from the large collection, and the experimental procedure was used to generate 64 346 term-date pairs. Diagrams corresponding to figures 1-5 with respect to the 64 346 term-date pairs are depicted in figures 6-10. Figures 1-4 show similar patterns to Figures 6-9. Figure 9, however, shows a slight increase in the proportion of agreements for the reduction levels 98% - 99%. There is also a significant difference between figures 5 and 10. Figure 5 shows that χ^2 and *I* agree with each other the most. In contrast, Figure 10 shows that χ^2 and *MI* agree with each other the most. Both figures, however, show that *MI* and *I* exhibit relatively strong disagreement.

4.2.4 Discussion

As stated in section 4.2.3, the proportion of agreements for *I* sharply decreases earlier than χ^2 and *MI* (89% for *I* and 96% for *MI* and χ^2). This agrees with other research (Yang & Pedersen, 1997; Sebastiani, 2002). *I* takes the presence and absence of terms into account. This results in a higher degree of term reduction than the other two methods. In their study, Yang & Pedersen (1997) indicate that this aggressive reduction results in a better classification performance, through effectiveness in excluding insignificant terms, whilst still keeping significant terms. In the current study, however, this may not be the case as some highly significant terms may be excluded because of the very high values that less genuinely significant terms may have.

With respect to the average proportion of miss rates for 39 reduction levels, *MI* has the largest proportion of miss rates (20.40%). Figure 3 shows that for reduction levels between 75% and 90%, *MI* tends to exclude a large proportion of terms judged significant (an average 26.20% miss rate), whilst χ^2 (0% miss rate) and *I* (an average 12.95% miss rate) do not. This finding suggests that *MI* is different from the other two methods in the way it ranks the significance of features. As discussed in section 4.2.2, however, it does not necessarily means that *MI* is less effective for aggressive term elimination than the other two methods.

As figures 5 and 10 show, *MI* and *I* disagree relatively strongly. This is due to the fact that *I* considers the amount of uncertainty in the presence and absence of a term, t_i, by observing the presence and absence of a date, d_j, as stated in the equations (4) and (5). In contrast, *MI* only considers the joint probability, $P(t_i, d_j)$, the probability that the term, t_i is present, $P(t_i)$, and the probability that the date, d_j is present, $P(d_j)$. Table 5 shows an example of (1) the usefulness of χ^2, *I* and *MI* values as indicators of the degree of term significance on a given date, and (2) an illustration of how *MI* behaves differently from χ2 and *I*. As listed in Table 4, the minimum threshold values of each method are χ^2=6.64, *I*=4.83E-6 and *MI*=3.93. Table 5 shows that *MI* starts to judge the term ‘asian nations’ to be significant on 12/10/2004, but, χ^2 and *I* on 26/12/2004, i.e. *before* the tsunami. The three methods agree that the term is significant between 26/12/2004-28/12/2004.

Overall, χ^2 selects the largest proportion of terms judged significant (i.e., in the 2-3 votes set). Thus, according to the evaluation method used, χ^2 is the best of the three methods. In addition, χ^2 tends to agree with both *MI* and *I*, with *MI* and *I* showing significant disagreement. Aside from the evaluation method used, it is possible that χ^2 might be over-fitting, i.e. tending to judge a term to be significant when it is not. This could occur due to the statistical problem that χ^2 is not reliable when the cell value is less than 5. In our experimental setting, the observed frequency, O(a) < 5 = 93.62%, O(b) < 5 = 0.0059%, O(c) < 5 = 68.39% and O(d) < 5 = 0%. This suggests that the χ^2 method would be unreliable for an overwhelming majority of the data, which does not sit comfortably with its evaluation success. This issue can be further investigated by analysing terms that were judged to be highly significant.

The basic assumption of using any kind of feature selection method is that insignificant terms can reliably be excluded by increasing threshold values. The experimental evidence, however, suggests otherwise. The three selection methods showed strong disagreement at a high level of reduction. *I* and *MI* disagree with each other the most, as they identify different types of unusual/significant behavior, whilst χ^2 identifies both types of unusual/significant behavior, but the key question is whether either or both of these types of unusual behavior represent genuinely significant real-world events. Table 6 shows an example of 5 terms which were assigned extremely large χ^2 values, along with their associated *I* and *MI* values. Excerpts of the RSS items posted on relevant dates are shown below.

[26/05/2004]: If you start with easy to grow *varieties* of plants and flowers, you can get *reliable results* without a lot of work.

[03/06/2004]: You can justify the slightly increased price when you think of the *extra work and expense* involved with just getting certified as organic.

[06/09/2004]: Building a pond was incredibly easy. Getting the *hole* dug was the *hardest part*.

The corpus used contained two significant world-events, i.e. the US election (3/11/2004) and the tsunami (26/12/2004). The χ^2, *I* and *MI* values of ‘US election’ and ‘tsunami’ are used for comparison. For the term, ‘US election’, the maximum value of χ^2, *MI* and *I* are 6066.67, 10.95 and 5.31E-05. For the term, 'tsunami' 420.81, 2.54 and 2.0E-4. It is conceivable that the five terms in Table 6 are highly significant on a particular date. It is, however, highly unlikely that they are much more significant than the two terms representing two world-events. This shows that the assumption that the larger a value, the higher the degree of the significance of a term is not generally true, even for the χ^2 method which has results that tend to corroborated by either *MI* or *I* (i.e. found in the 2-3 votes set).

5. Conclusions

The three feature selection methods, χ^2, Mutual Information (*MI*) and Information Gain (*I*), were evaluated in terms of their judgement on determining the degree of the significance of a term on a certain date. The evaluation method used pointed to χ^2 as being the best of the three methods. Nevertheless, it is far from perfect as the examples of extremely high values assigned to relatively insignificant terms showed (Table 6). As a corollary of this, the hypothesis that large values reliably indicate highly significant terms was not supported. The experimental evidence suggests that the three methods will have a significant degree of disagreement for some terms assigned extremely large values.

Finally, the voting method used to compare the three methods provided a great deal of illumination on the differences between the methods and was able to suggest a best method. Ultimately, however, human classifications might be needed to decide which method was the best at identifying significant terms at different reduction levels, because even extremely high values were not reliable indicators of significance. Nevertheless, since the evaluation method showed that there is no convergence between methods for high levels of reduction, it follows that caution would always be needed in interpreting results, whichever method is used. It may be that heuristics such as upper thresholds or the exclusion of low frequency terms might be required to provide improved results but a full human classification exercise would be needed to evaluate such methods.

Acknowledgements

The work was supported by a European Union grant for activity code NEST-2003-Path-1. It is part of the CREEN project (Critical Events in Evolving Networks, contract 012684). We thank the reviewers for their helpful comments.

Figures

Legend for figures 1-4 and 6-9:

Δ : the proportion of agreements; ∇ : the proportion of disagreements; × : miss rate.

Legend for figures 5 and 10:

◇: the proportion of agreements between χ_2 and *I*; ○: the proportion of agreements between χ_2 and *MI*; □: the proportion of agreements between *MI* and *I*.

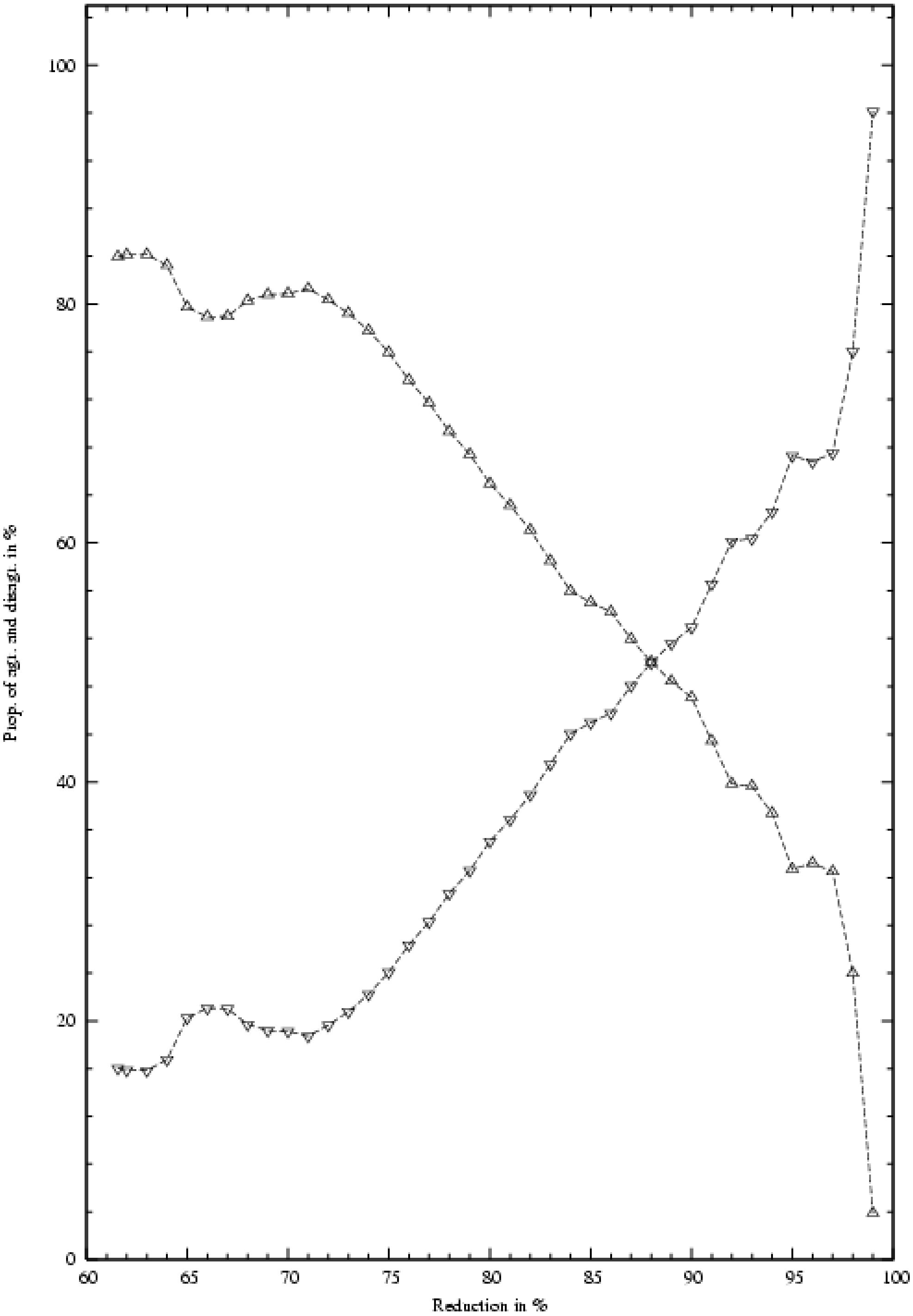

Fig 1: The proportion of agreements and disagreements between χ^2, *MI*, *I* and the 3-votes set.

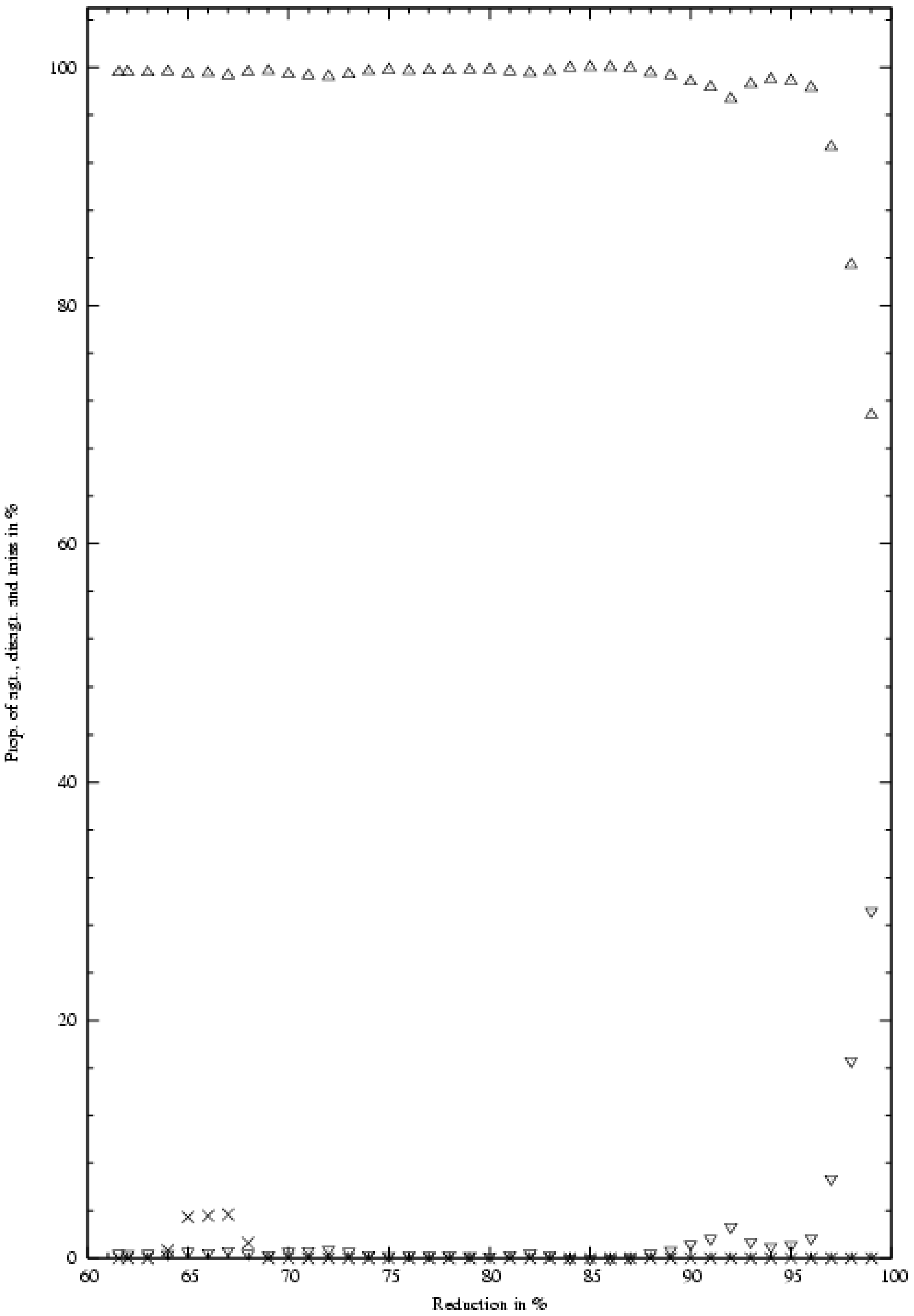

Fig 2: The proportion of agreements, disagreements and miss rates between χ^2 and the 2-3-votes set.

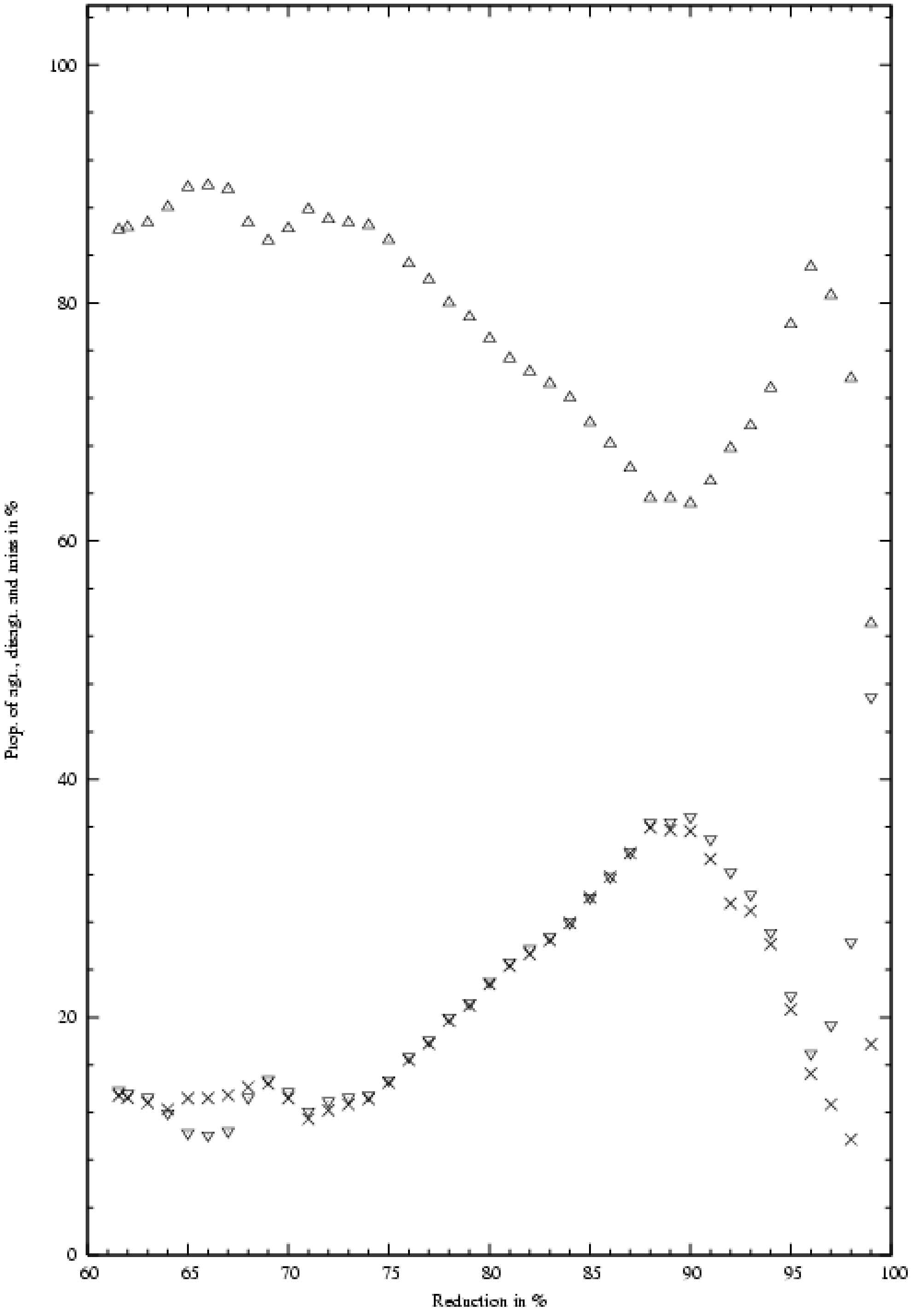

Fig 3: The proportion of agreements, disagreements and miss rates between *MI* and the 2-3-votes set.

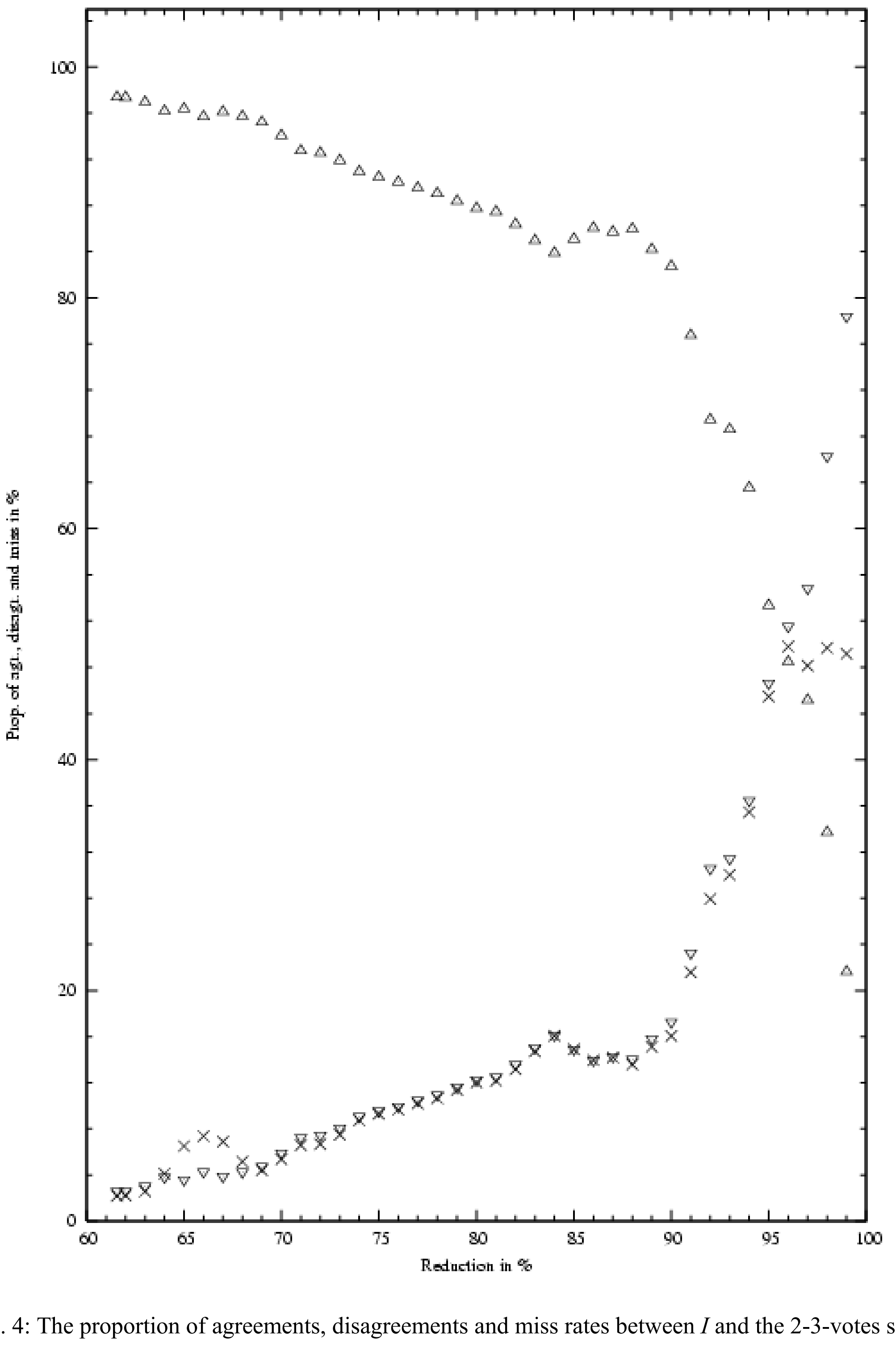

Fig. 4: The proportion of agreements, disagreements and miss rates between *I* and the 2-3-votes set.

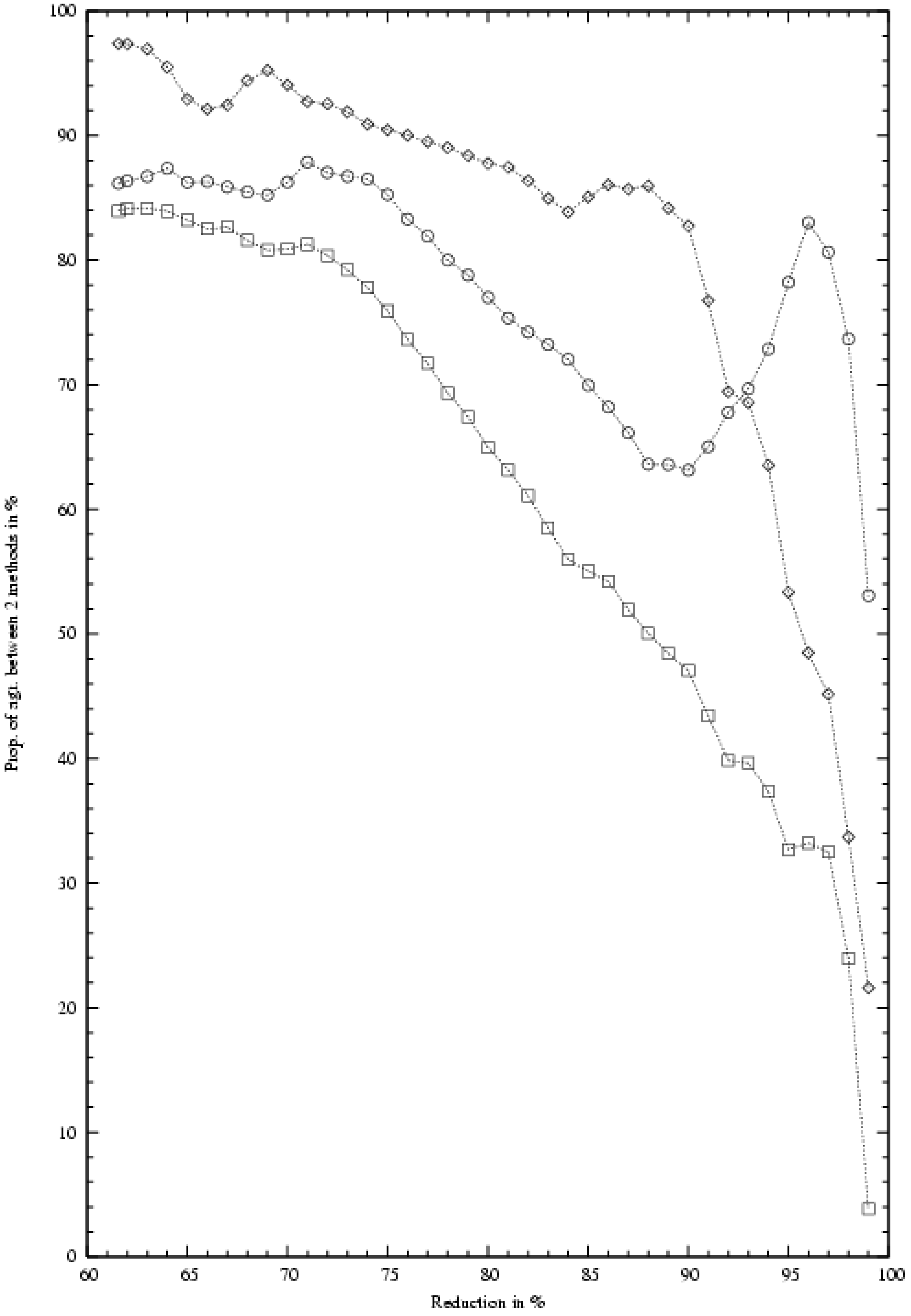

Fig. 5: The proportion of agreements between two methods.

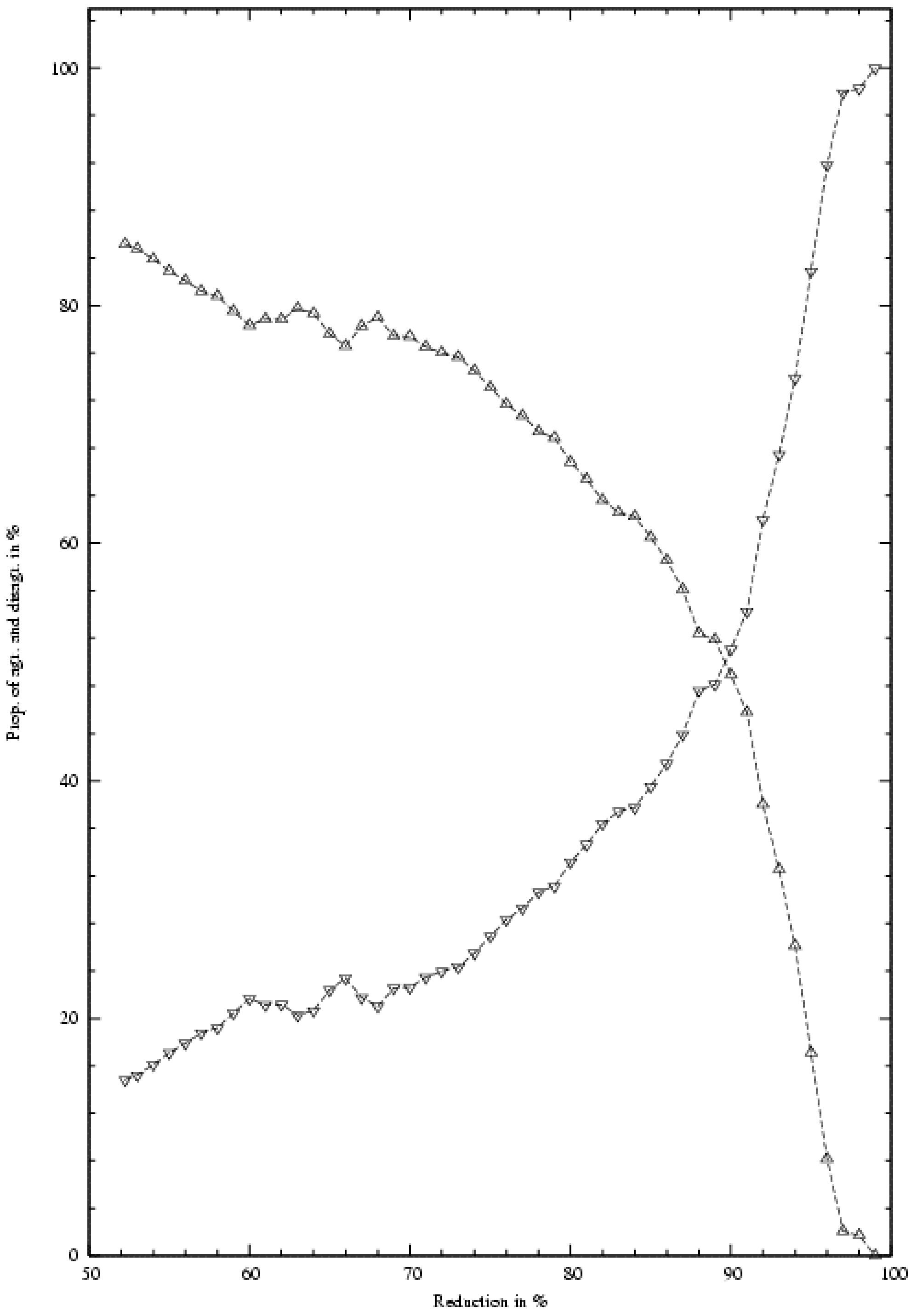

Fig 6: The proportion of agreements and disagreements between χ^2, *MI*, *I* and the 3-votes set.

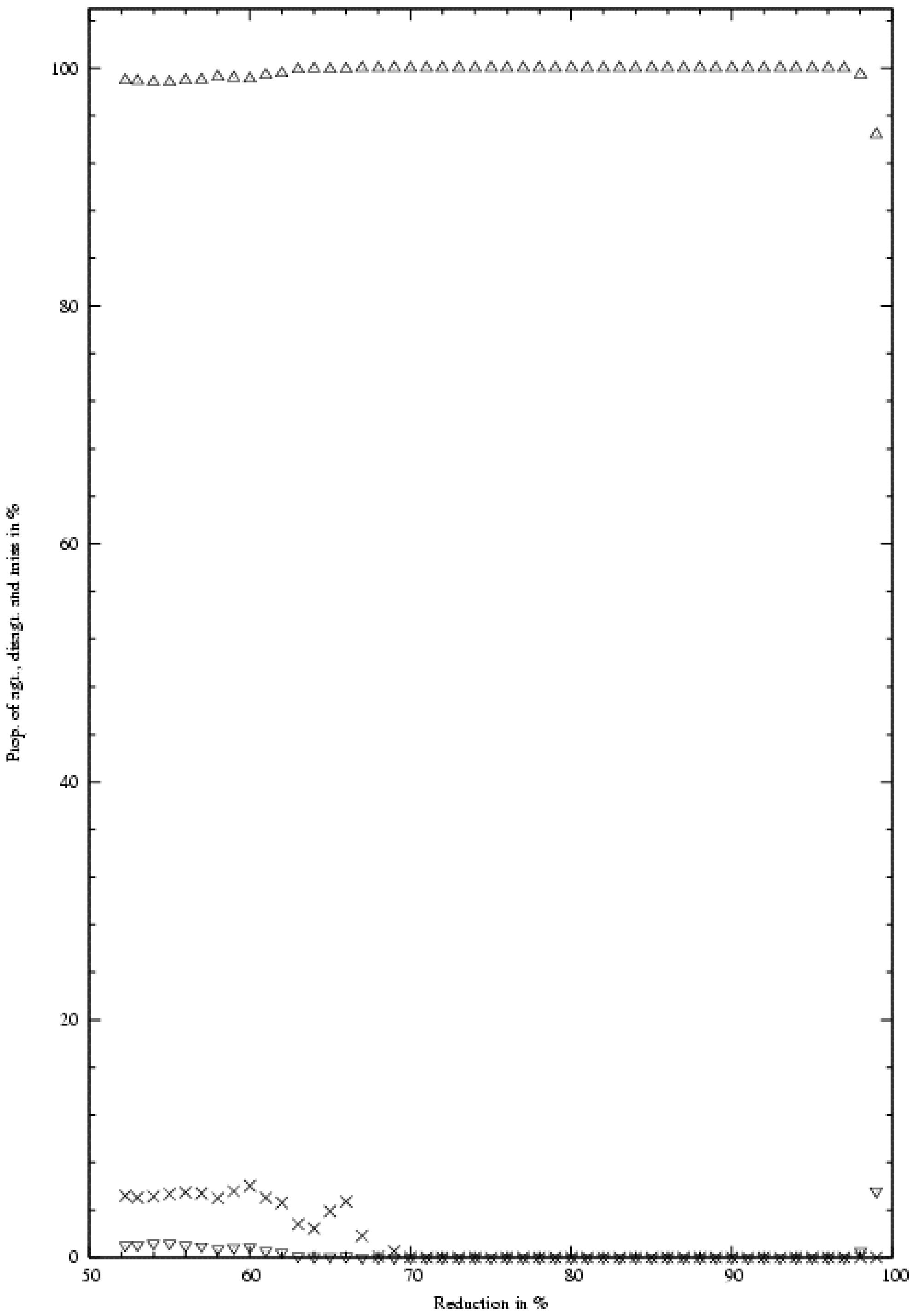

Fig 7: The proportion of agreements, disagreements and miss rates between χ^2 and the 2-3-votes set.

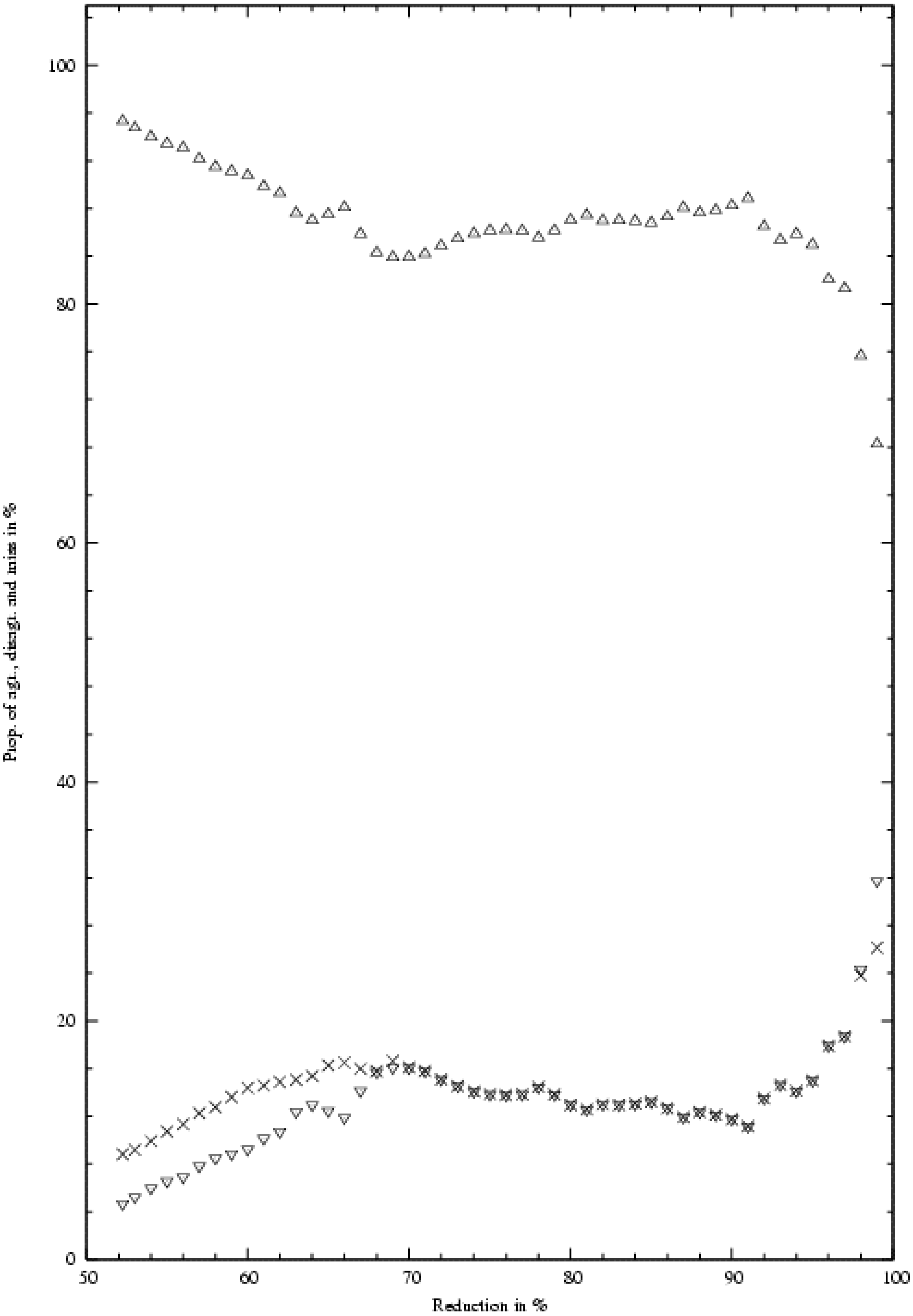

Fig 8: The proportion of agreements, disagreements and miss rates between *MI* and the 2-3-votes set.

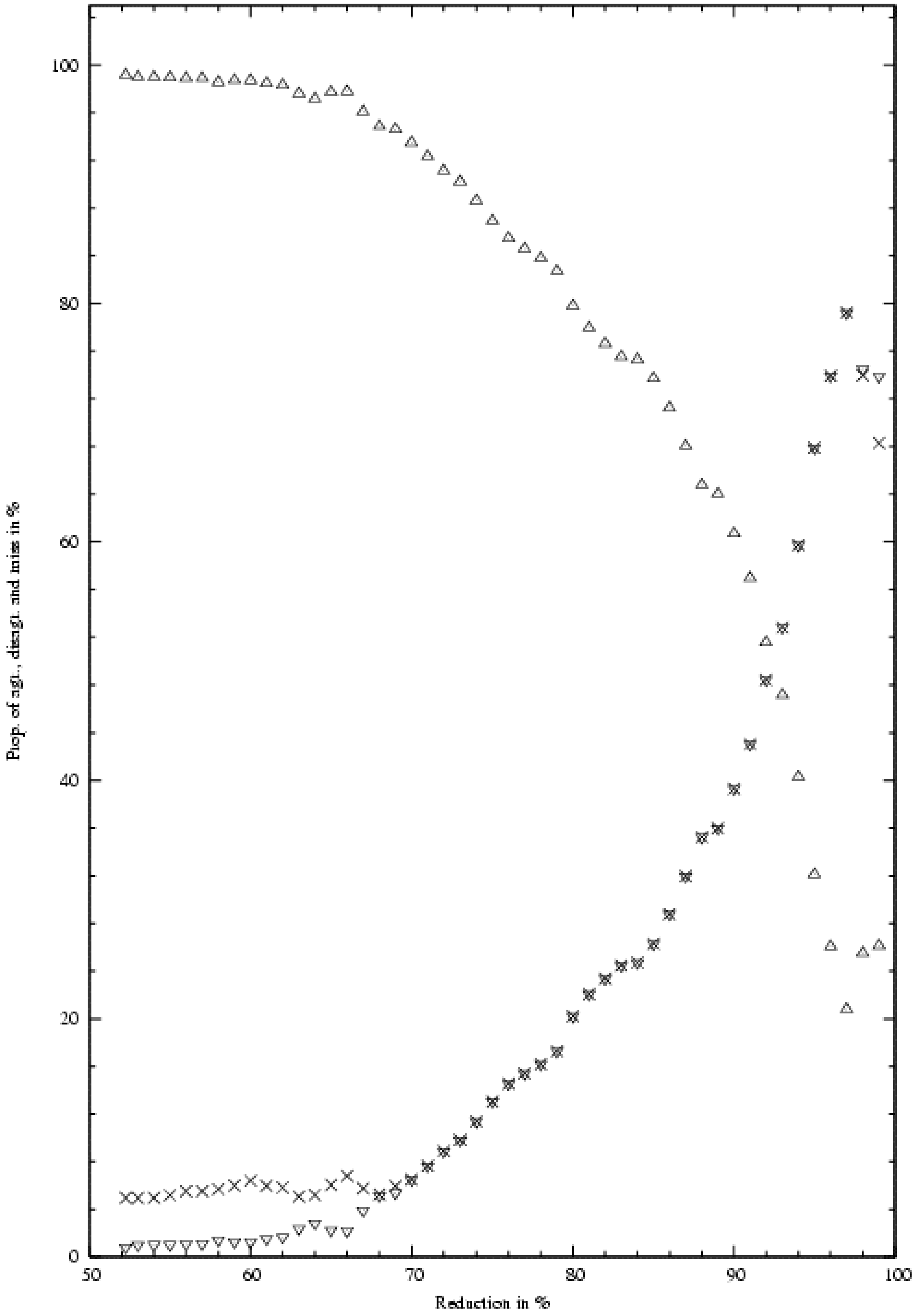

Fig. 9: The proportion of agreements, disagreements and miss rates between *I* and the 2-3-votes set.

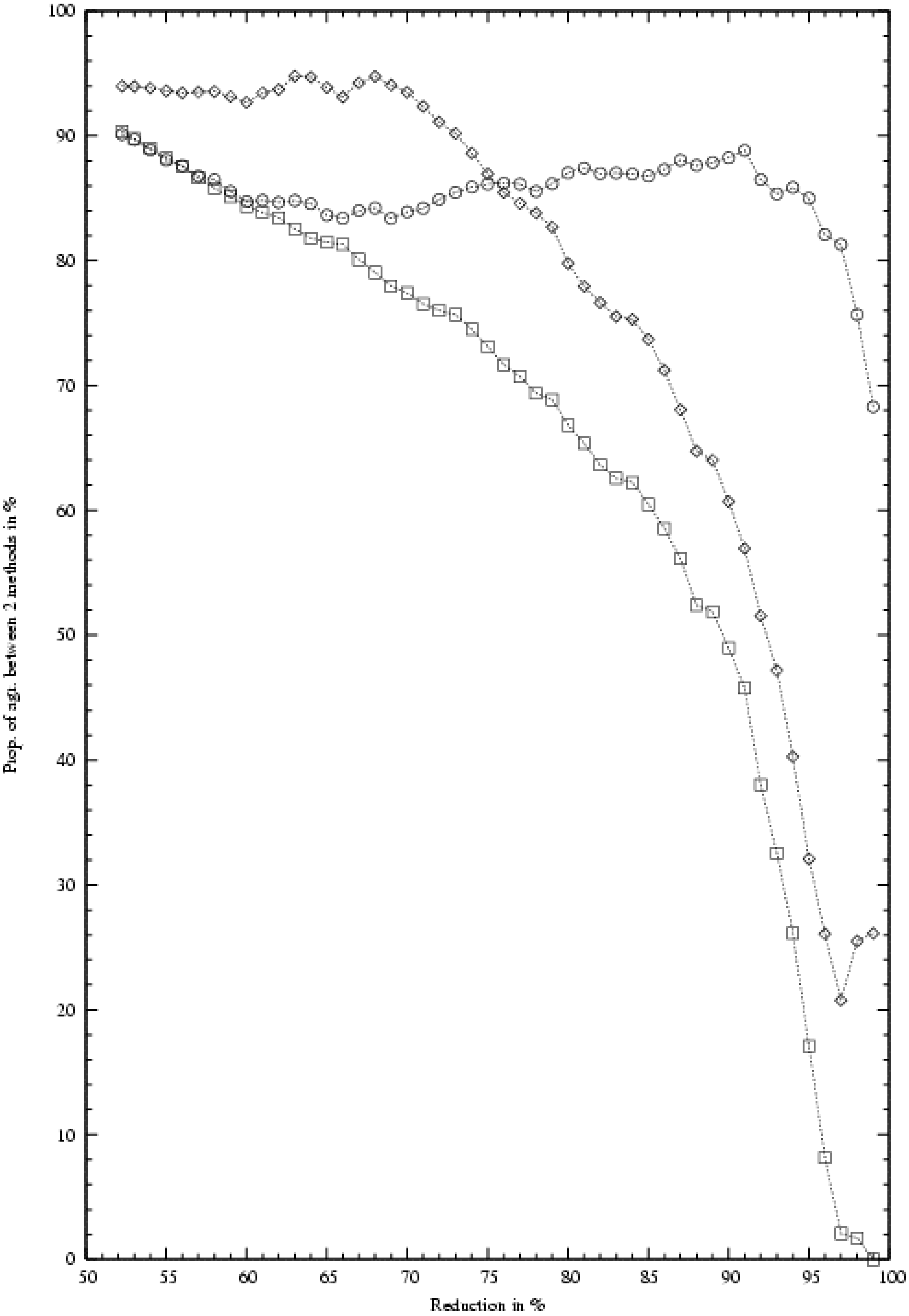

Fig. 10: The proportion of agreements between two methods.

Tables.

Topic Detection	Generation of Overview Timelines
For each story, assign each term a weight from a weighting function. Set a threshold with respect to weight values. Select significant terms as the content representation of a story. Apply a clustering algorithm to cluster stories that are topically related. Each cluster contains a number of stories that represent a topic.	For all stories, assign each term a value with respect to a certain time period, by using χ^2 statistics. Set a threshold with respect to χ^2 values. Select significant terms for a certain time period. Apply a clustering algorithm to cluster significant terms that are closely related. Each cluster contains a number of significant terms that represent a topic.

Table 1: The difference in sequential processes between topic detection and generation of overview timelines.

Topic Detection	
Strengths	Weaknesses
It can cluster stories that discuss the same topic. This can be used to generate a summary of the stories. Given a new, incoming story, the significant terms found within the new story can be used to determine whether the new story discusses the previously identified topic or a new topic (on-line detection).	It is computationally expensive to cluster a collection of stories. This is especially significant if there is a large collection of stories. When a new topic is detected, the weights of all the existing terms need to be recalculated since the weight values depend on the total number of stories in the corpus.

Table 2: The strengths and the weaknesses of the topic detection approach.

Generation of Overview Timelines	
Strengths	Weaknesses
It can cluster significant terms within a certain time period to identify their associated topics. The computational cost for the clustering is low, since only few significant terms on a certain time period are used. It is less complex than the topic detection approach, since the model only focuses on the generation of 2x2 contingency tables to compute the χ^2 value of each term for a given time period.	It is not designed for generating a summary of all the stories that are topically related. It is designed in a retrospective way, in the sense that it only analyses a corpus that contains a collection of stories. It does not have the necessary data to carry out online detection, i.e. to determine whether a new, incoming story discusses the previously identified topic or a new topic.

Table 3: The strengths and the weaknesses of the generation of overview timelines approach.

Reduction (%)	$\chi2$	*MI*	*I*	Reduction (%)	$\chi2$	*MI*	*I*
61.56	6.64	3.93	4.83E-6	81	26.35	6.47	8.61E-6
62	7.02	3.93	4.94E-6	82	28.02	6.52	8.87E-6
63	7.82	3.93	5.26E-6	83	31.07	6.54	9.08E-6
64	8.69	3.95	5.47E-6	84	36.99	6.57	9.28E-6
65	8.72	4.20	5.64E-6	85	39.78	6.64	9.38E-6
66	9.73	4.42	5.90E-6	86	44.29	6.69	9.65E-6
67	10.63	4.64	6.27E-6	87	49.48	6.76	1.03E-5
68	11.89	4.87	6.35E-6	88	55.26	6.85	1.11E-5
69	13.32	5.05	6.59E-6	89	70.53	6.99	1.21E-5
70	14.51	5.24	7.05E-6	90	87.88	7.27	1.33E-5
71	15.74	5.24	7.41E-6	91	102.83	7.37	1.43E-5
72	16.53	5.41	7.61E-6	92	116.92	7.51	1.55E-5
73	18.67	5.54	7.78E-6	93	154.60	7.65	1.59E-5
74	20.44	5.69	7.84E-6	94	202.87	7.95	1.66E-5
75	21.18	5.86	7.87E-6	95	288.30	8.49	1.79E-5
76	21.47	6.04	7.93E-6	96	487.76	9.37	2.02E-5
77	22.46	6.17	8.04E-6	97	827.17	10.92	2.45E-5
78	22.93	6.29	8.10E-6	98	1839.77	11.79	3.02E-5
79	24.02	6.41	8.20E-6	99	4049.16	13.08	4.07E-5
80	25.08	6.45	8.31E-6				

Table 4: The 39 levels of reduction along with their associated threshold values.

$\chi2$	*I*	*MI*	Date
4.06	3.60e-06	4.32	2004-10-12
46.20	1.20e-05	4.61	2004-12-26
287.60	3.71e-05	5.63	2004-12-27
34.70	1.07e-05	4.24	2004-12-28

Table 5: An excerpt of the $\chi2$, *I*, *MI* values, assigned to the term, 'asian nations', which are sorted according to date.

Term	$\chi2$	*I*	*MI*	Date
reliable results	696 530	5.3e-3	10.88	2004-05-26
varieties	563 869	4.93e-3	10.57	2004-05-26
extra work and expense	696 530	5.32e-3	10.88	2004-06-03
hardest part	620 846	5.08e-3	10.71	2004-09-06
hole	293 551	4.26e-3	9.60	2004-09-06

Table 6: An example of 5 outliers which are assigned an extremely large χ^2 value, along with their associated *I* and *MI* values.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998a). Topic detection and tracking pilot study: final report. In Proceedings of the DARPA broadcast news transcription and understanding workshop, February 8-11, 1998. Lansdowne, Virginia, USA.

Allan, J., Papka, R., & Lavrenko, V. (1998b). On-line new event detection and tracking. In Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, August 24-28, 1998 (pp. 37-45). Melbourne, Australia.

Avesani, P., Cova, M., Hayes, C., & Massa, P. (2005). Learning contextualised weblog topics. In Proceedings of the 14th international WWW conference: 2nd annual workshop on weblogging ecosystem: aggregation, analysis and dynamics, May 10, 2005. Chiba, Japan.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. In Proceedings of the 12th national conference on artificial intelligence (AAAI 1994), July 31 - August 4, 1994 (pp. 722-727). Seattle, Washington.

Chen, H., Hsu, P., Orwig, R., Hoopes, L., & Nunamaker, J. F. (1994). Automatic concept classification of text from electronic meetings. ACM Journal, 37(10), 56-73.

Church, K. W. & Hanks, P. (1989). Word association norms, mutual information and lexicography. In Proceedings of the 27th annual meeting of the Association for Computational Linguistics (ACL), June 26--29, 1989 (pp. 76--83). Vancouver, B.C.

Conrad, J. G. & Utt, M. H. (1994). A sytem for discovering relationships by feature extraction from Text Databases. In W. B. Croft & C. J. van Rijsbergen (Eds.), Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, July 3--6, 1994 (pp. 260--270). Dublin, Ireland.

Cooper, W. S. (1995). Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. ACM Transactions on Information Systems, 13(1), 100-111.

Dumais, S. & Chen, H. (2000). Hierarchical classification of Web content. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong, & P. Ingwersen (Eds.), Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, July 24-28, 2000 (pp. 256-263). Athens, Greece.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1),61-74.

Egghe, L. (2000). New informetric aspects of the Internet: some reflections - many problems. *Journal of Information Science, 26*(5), 329-335.

Erdös, P. & Rényi, A (1960). On the evolution of random graphs. Publications of the mathematical institute of the hungarian academy of science, 5(1), 17-61.

Gill, K. E. (2004). How can we measure the influence of the blogosphere? In Proceedings of the 13th international WWW conference: workshop on weblogging ecosystem: aggregation, analysis and dynamics, May 18, 2004. New York, USA.

Gill, K. E. (2005). Blogging, RSS and the information landscape: a look at online news. In Proceedings of the 14th international WWW conference: 2nd annual workshop on weblogging ecosystem: aggregation, analysis and dynamics, May 10, 2005.Chiba, Japan.

Glance, N. S., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: automated trend discovery for weblogs. In Proceedings of the 13th international WWW conference: workshop on weblogging ecosystem: aggregation, analysis and dynamics, May 18,2004. New York, USA.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through Blogspace. In Proceedings of the 13th international WWW conference, May 17-22, 2004 (pp. 491—501). New York, USA.

Hammersley, B. (2005). Developing feeds with RSS and Atom. Sebastopol, CA: O'Reilly.

Jenkins, C., Jackson, M., Burden, P., & Wallis, J. (1999). Automatic RDF metadata generation for resource discovery. In Proceedings of the 8th international WWW conference, May 11-14, 1999 (pp. 227-242). Toronto, Canada.

Jin, H., Schwartz, R., Sista, S., & Walls, F. (1999). Topic tracking for radio, tv broadcast, and newswire. In Proceedings of the DARPA broadcast news workshop, February 28 - March 3, 1999. Herndon, Virginia, USA.

Kwon, O.-W. & Lee, J.-H. (2000). Web page classification based on k-nearest neighbor approach. In Proceedings of the 5th international workshop on information retrieval with Asian languages, September 30 - October 1, 2000 (pp. 9-15). Hongkong, China.

Larkey, L. S. (1998). Some issues in the automatic classification of U.S. patents. In Proceedings of the 15th national conference on artificial intelligence (AAAI 1998): workshop on learning for text categorization, July 26 - 30, 1998. Madison, Wisconsin, USA.

Luo, X. (2004). Information extraction for new event detection. IBM. http://www.nist.gov/speech/tests/tdt/tdt2004/papers/IBM-NED-TDT2004.ppt accessed November 28, 2005.

MacKay, D. J. C. (2003). Information theory, inference, and learning algorithms. Cambridge University Press, 2nd edition.

Moore, A. (2003). Information Gain. <http://www-2.cs.cmu.edu/ awm/tutorials/infogain.html> (accessed 3 January 2005).

Ng, H. T., Goh, W. B., & Low, K. L. (1997). Feature selection, perception learning, and a usability case study for text categorization. In N. J. Belkin, A. D. Narasimhalu, P. Willett, W. Hersh, F. Can, & E. Voorhees (Eds.), Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval, July 27-31, 1997 (pp. 67-73). Philadelphia, Pennsylvania, USA.

NIST Speech Group (2005). The topic detection and tracking phase 2 (TDT2) evaluation plan. NIST. < http://www.nist.gov/speech/tests/tdt/tdt98/ > (accessed 15 June 2005).

Pikas, C. K. (2005). Blog searching for competitive intelligence, brand image, and reputation management. Online, 29(4), 16-21.

Ramshaw, L. A. & Marcus, M. P. (1995). Text chunking using transformation-based learning. In D. Yarovsky & K. Church (Eds.), Proceedings of the 3rd workshop on very large corpora (VLC 1995), 1995 (pp. 82-94). Massachusetts, USA.

Schultz, J. M. & Liberman, M. (1999). Topic detection and tracking using idf-weighted cosine coefficient. In Proceedings of the DARPA broadcast news workshop, February 28 - March 3, 1999. Herndon, Virginia, USA.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1-47.

Shannon, C. E. & Weaver, W. (1963). The mathematical theory of communication. University of Illinois Press.

Swan, R. & Allan, J. (1999). Extracting significant time varying features from text. In S. Gauch & I.-Y. Soong (Eds.), Proceedings of the 8th international conference on information and knowledge management (CIKM 1999), November, 1999 (pp. 38-45). Kansas City, USA.

Swan, R. & Allan, J. (2000). Automatic generation of overview timelines. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong, & P. Ingwersen (Eds.), Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, July 24-28, 2000 (pp. 49-56). Athens, Greece.

Swan, R. & Jensen, D. (2000). TimeMines: constructing timelines with statistical models of word usage. In Proceedings of the 6th ACM SIGKDD conference on knowledge discovery and data mining: workshop on text mining, August 20-23, 2000 (pp. 73-80). Boston, MA, USA.

Thelwall, M., Prabowo, R. & Fairclough, R. (2006, to appear). Are raw RSS feeds suitable for broad issue scanning? A science concern case study. Journal of the American Society for Information Science and Technology.

Walls, F., Jin, H., Sista, S., & Schwartz, R. (1999). Topic detection in broadcast news. In Proceedings of the DARPA broadcast news workshop, February 28 - March 3, 1999. Herndon, Virginia, USA.

Yang, Y. (1994). Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In W. B. Croft & C. J. van Rijsbergen (Eds.), Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, July 3-6, 1994 (pp. 13-22). Dublin, Ireland.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Information Retrieval, 1(1-2), 69-90.

Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In Proceedings of the 14th international conference on machine learning (ICML 1997), July 8-12, 1997 (pp. 412-420). Nashville, Tennessee.

Yang, Y., Pierce, T., & Carbonell, J. (1998). A study on retrospective and on-line event detection. In Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, August 24-28, 1998 (pp. 28-36). Melbourne, Australia.