Comparisons of the Cosine Measure and Sub-String Indexing on Usenet News Articles

Fredrik Kilander
fk@dsv.su.se

January 4, 1995

Abstract

Two document similarity measures are compared on a sample of Usenet News messages.

1 Introduction

In order to estimate the similarity between a query and a collection of documents in a database, a number of measures has been applied and tried in the area of information retrieval. For the purpose of a filtering application, these devices appear to be good starting points in the selection of one of more measures that are to provide a numerical distance between unread messages and a set of long-term queries established by the user. This paper describes a set of evaluations applied to two such similarity measures, the cosine measure and sub-string indexing.

2 The Cosine Measure

The cosine measure [SM83] has received its name from a trigonometrical analogy: the angle between two vectors in N-dimensional space is always measurable across a 2-D plane. When the vectors point in the same direction the angle between them is $\phi = 0$. When they are perpendicular to each other the angle between them is $\phi = \pi/2$. Taking the cosine of the angle normalizes the range between the extremes into

$$\{\cos(\phi = \pi/2) = 0 \ldots \cos(\phi = 0) = 1\} .$$

In this *vector-space* model, the desired result is directly available by taking the cross product of the vectors. For texts, each document or query is represented by a vector of weighted terms:

$$V = \{wt_0, wt_1, \ldots, wt_n\}$$

where each wt_i consists of two parts, the numerical weight and the term. The term is some component feature of the document, usually a string containing

a word or a phrase, e. g. `computer` or `concept graphs`. In addition, the term may be any boolean feature of the document.

Given a document d and a query q, the cosine similarity was computed in the following way:

$$cosine_{q,d} = \frac{\sum_t w_{q,t} \cdot w_{d,t}}{\sqrt{\sum_t w_{q,t}^2} \cdot \sqrt{\sum_t w_{d,t}^2}}$$

where $w_{x,t}$ was the number of times term t occurred in document or query x. The *Inverse Document Frequency* has been implemented but was not used in these experiments since each query and document only contained a single article.

3 Sub-String Indexing

This similarity measure (Morita and Shinoda [MS94]) is simply the proportion of common terms in the document and the query:

$$sim = \frac{2 \mid V_q \cap V_d \mid}{\mid V_q \mid + \mid V_d \mid}.$$

4 Experiments

A sample of articles consisting of 101 articles was taken in non-sequential order from a newsgroup with a number of ongoing threads and heterogeneous postings (alt.games.doom). One article was selected to act as the query. All articles in the sample (including the query itself) were then compared against the query and their similarity registered. The message headers were included in the comparison. The average size of the articles (including the message headers)[1] were 34.07 lines (std. dev. 21.12), 177.46 words, (std. dev. 134.79) and 1381.66 characters (std. dev. 868.42).

Two dimensions were tested for each similarity measure: terms constructed from single words and two-word phrases created by combining adjacent pairs of words. Digits and alphabetics were considered to be part of words. Other characters were considered to be word delimiters. The other tested dimension was the case of alphabetic characters. In the following text the phrase *case* means that the case of alphabetic characters was significant in comparison and *no case* means that it was not significant. No stop-list of common words was employed.

4.1 Allowing Case to be Significant

Figures 1, 2, 3 and 4 shows the similarity of the two measures using a single word (figure 1 and 2) and two word phrases (figure 3 and 4).

[1] As reported by the UNIX utility *wc*.

Three distinct features should be noticed: the narrowness of the filter, the slope of the flattened curve and the average level of the flattened curve. All three influence decisions regarding thresholds for inclusion or exclusion.

The first point in all diagrams is the query message. Since it matches itself perfectly it is assigned a score of 1. The cosine measure, for one and two word phrases distinctly indicates two articles closely related to the query, while the sub-string index only manages to indicate one. The sub-string index creates a more narrow filter than the cosine measure and has a flatter slope over the non-relevant portion. The two word phrases apparently increases discrimination and lowers the average level of the slope (implicitly flattening it).

Figure 5 shows all four measures combined. Note that the article ordering is uncorrelated between the graphs. The figure shows the difference in slope between cosine and sub-string index and also the general effect of two word phrases.

4.2 Ignoring Case

Figures 6, 7, 8 and 9 show the result of repeating the previous experiment with case being insignificant. The has the result of lowering the number of unique words in the articles and increasing the frequency of some words; fewer words that are repeated more often. The expected result is of course that documents become more similar to each other and this is reflected in the graphs. Otherwise the behaviour of the similarity measures are basically the same as with case sensitivity. Figure 10 collapses the four graphs into one figure. Article ordering is uncorrelated. Compared with figure 5 where case was significant, the intersection between cosine and sub-string index for 1 word has moved to the right, mostly because the average level of the cosine measure increased when the case distinction was removed.

4.3 Case Difference—Cosine One Word

Figure 11 show the cosine measure with and without case significance. Article comparisons are sorted according to the significant case measurement. In all cases, the same test without case distinction receives an equal or higher score (more similar). The difference between scores for a particular article can be explained by the amount of dissimilarity contained in words distinguished by case only.

4.4 Case Difference—Sub-String One Word

Figure 12 show the sub-string index measure with and without case significance. Article comparisons are sorted according to the significant case measurement. In all cases, the same test without case distinction receives an equal or higher score (more similar). The difference between scores for a particular article can be explained by the amount of dissimilarity contained in words distinguished by case only.

	Words	Case	No case
Cosine	1	0.264948 (0.115154)	0.307219 (0.117363)
	2	0.073103 (0.114970)	0.073977 (0.115743)
Substring	1	0.259642 (0.086204)	0.282888 (0.084398)
	2	0.073875 (0.098067)	0.077426 (0.097920)

Table 1: Average similarities (and standard deviation).

4.5 Case Difference—Cosine Two Words

Figure 13 shows the cosine on phrases of two words. With the exception of the second article (at a similarity level close to 0.61) the differences are very small. A smaller region of the similarity range is shown in figure 14. In this figure, both axis plot similarity. The horizontal axis is for the case significant measurement and the vertical axis for the data set where case was not significant. A line has been drawn across the graph—articles below the line receieved a greater score when case was important; articles above the line where more similar to the query when case was not important.

As can be seen in figure 14, a number of articles received a *lower* score when case differences were removed. The reason for this is not immediately clear, but is likely found in the difference in vector lengths; the term vectors are longer when case is distinct and therefore encode more dimensions.

4.6 Case Difference—Sub-String Two Words

Figure 15 shows the sub-string index on phrases of two words. With the exception of the second article (at a similarity level close to 0.34) the differences are very small. A smaller region of the similarity range is shown in figure 16. In this figure, both axis plot similarity. The horizontal axis is for the case significant measurement and the vertical axis for the data set where case was not significant. A line has been drawn across the graph—articles below the line receieved a greater score when case was important; articles above the line where more similar to the query when case was not important.

Here all articles receive a greater similarity score relative to the query when the case distinction was removed. This is not surprising, since the relief results in a lower number of unique phrases.

4.7 Statistics

Table 1 summarizes the average measurements for the eight cases reviewed here. In summary, the cosine measure on single words gave the highest average similarity, and also the lowest for two word phrases under case significance.

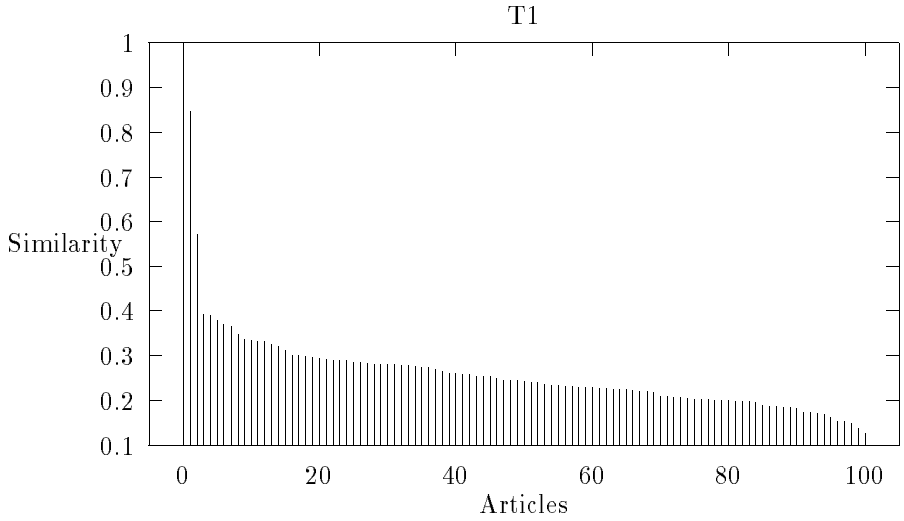

Figure 1: Cosine, case, one word.

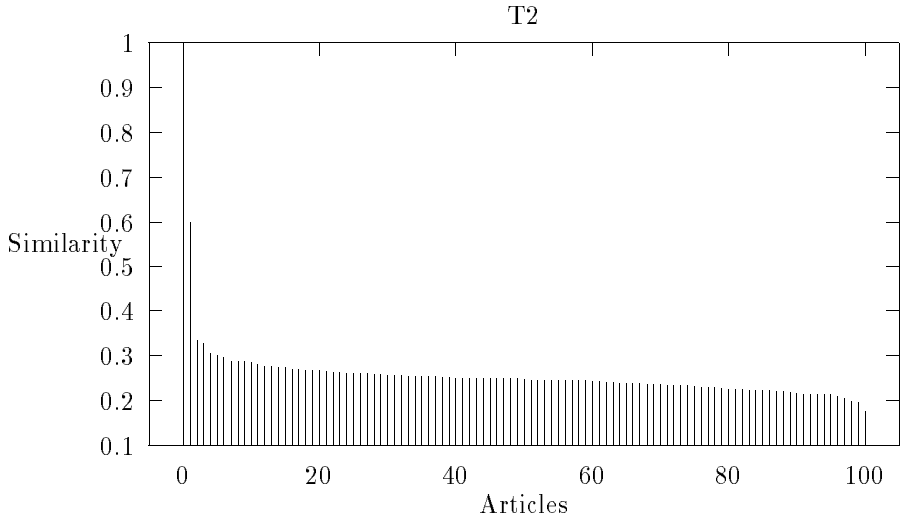

Figure 2: Sub-string, case, one word.

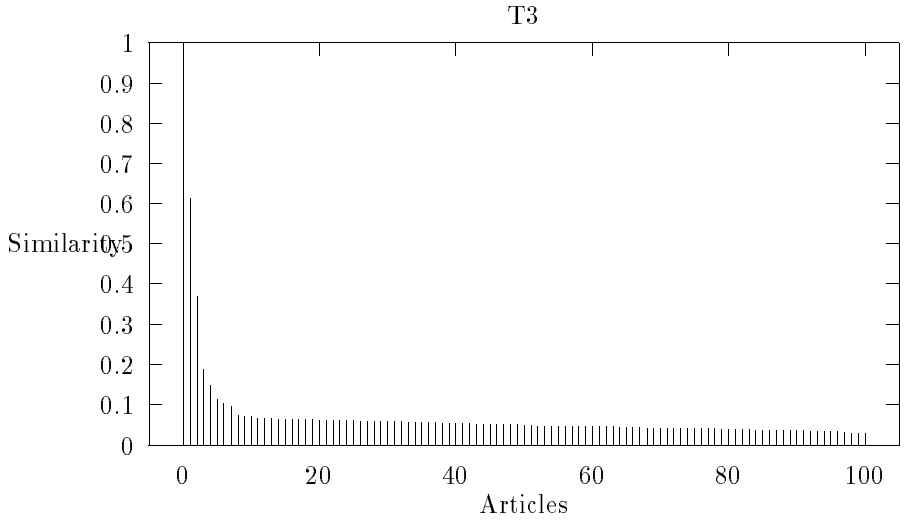

Figure 3: Cosine, case, two word phrase.

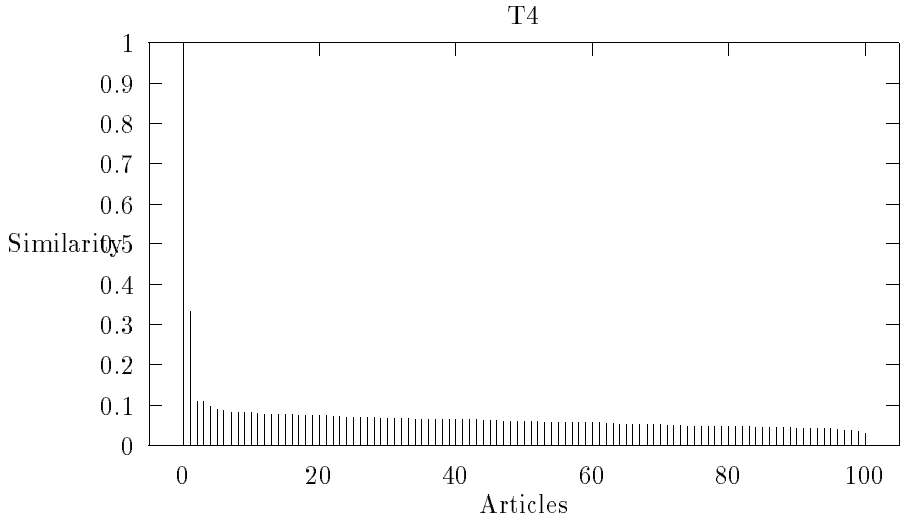

Figure 4: Sub-string, case, two word phrase.

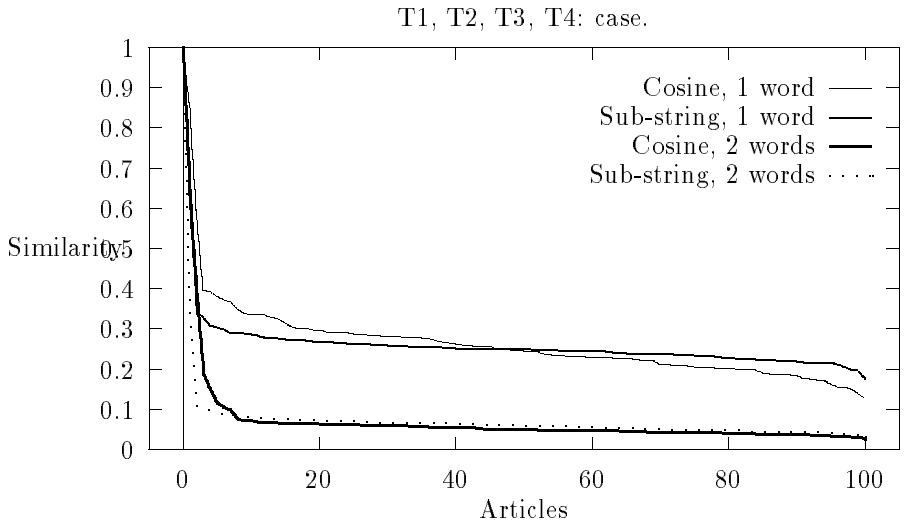

Figure 5: Cosine, sub-string, case, one and two word phrases. Article ordering is uncorrelated.

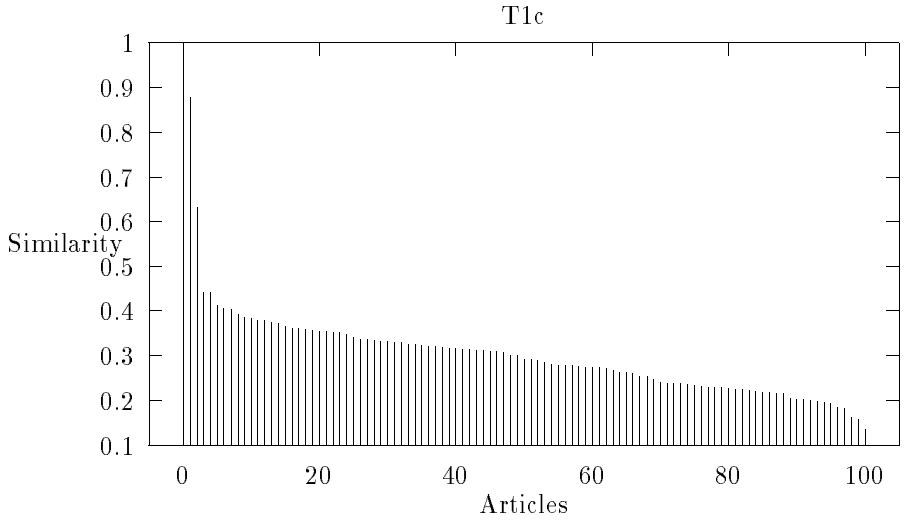

Figure 6: Cosine, no case, one word.

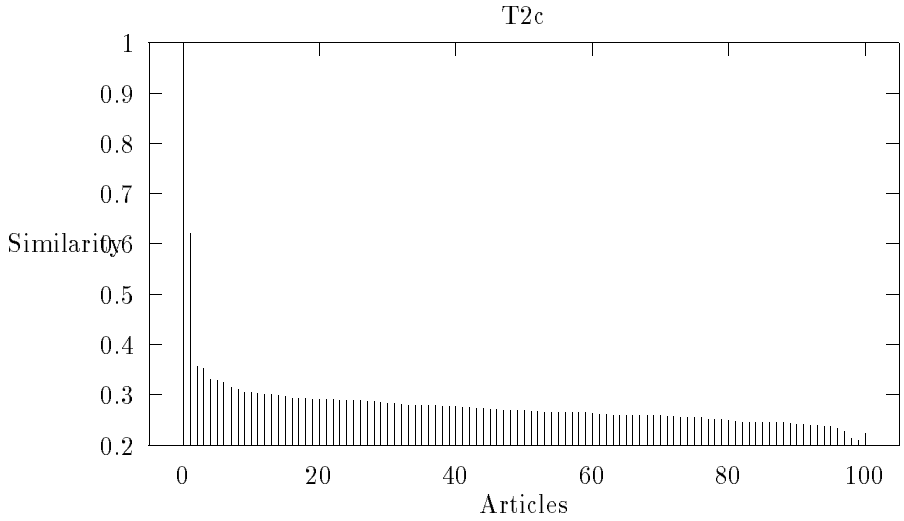

Figure 7: Sub-string, no case, one word.

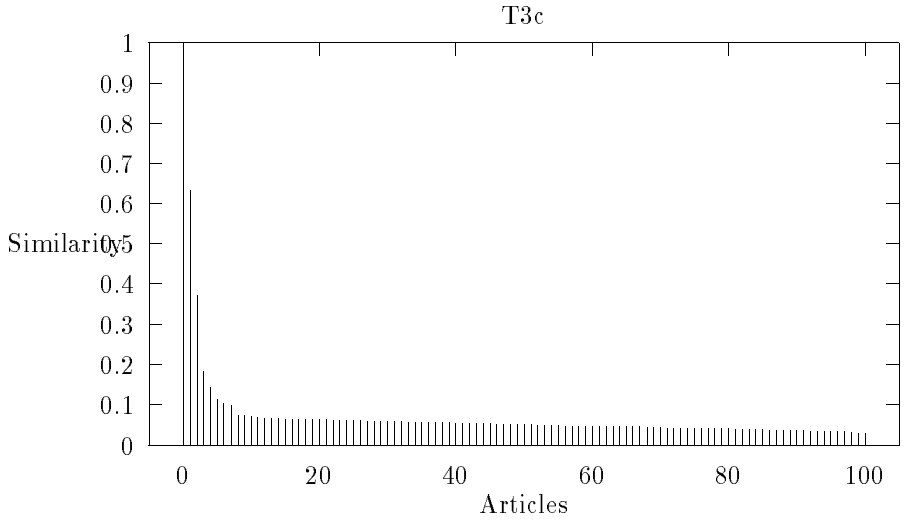

Figure 8: Cosine, no case, two word phrase.

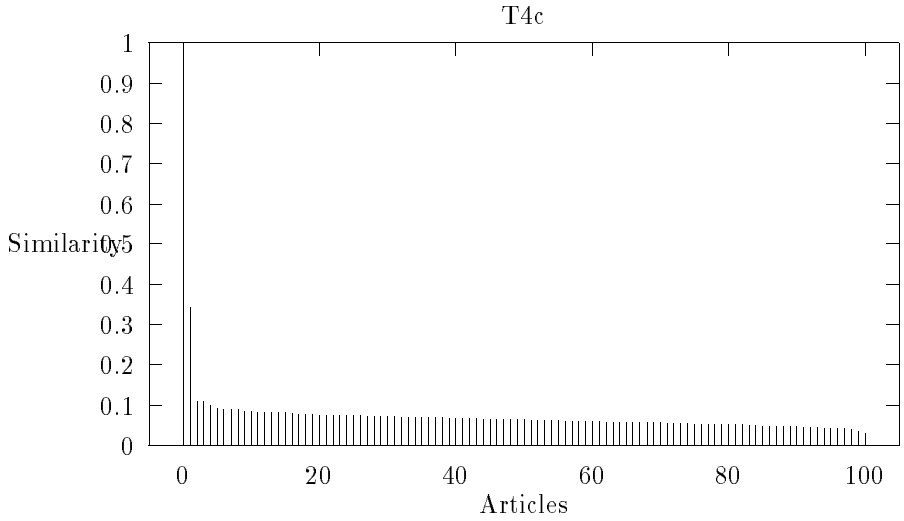

Figure 9: Sub-string, no case, two word phrase.

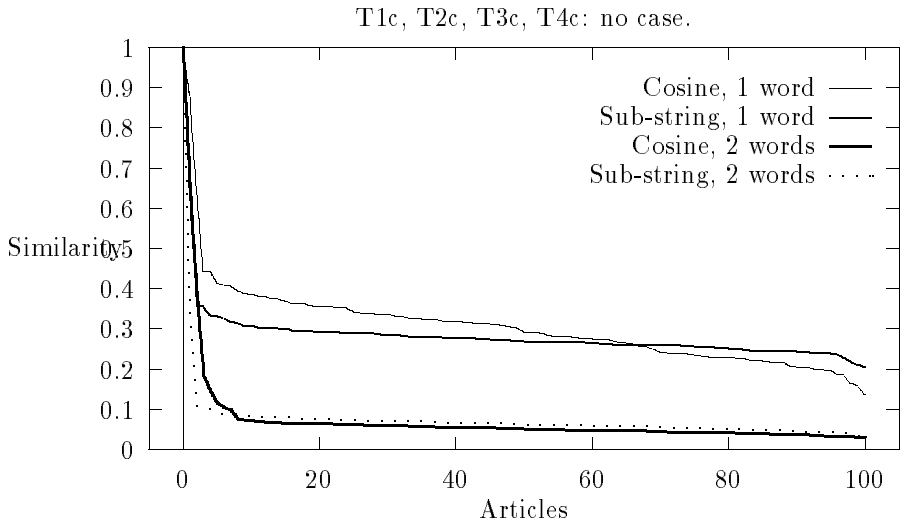

Figure 10: Cosine, sub-string, no case, one and two word phrases. Article ordering is uncorrelated.

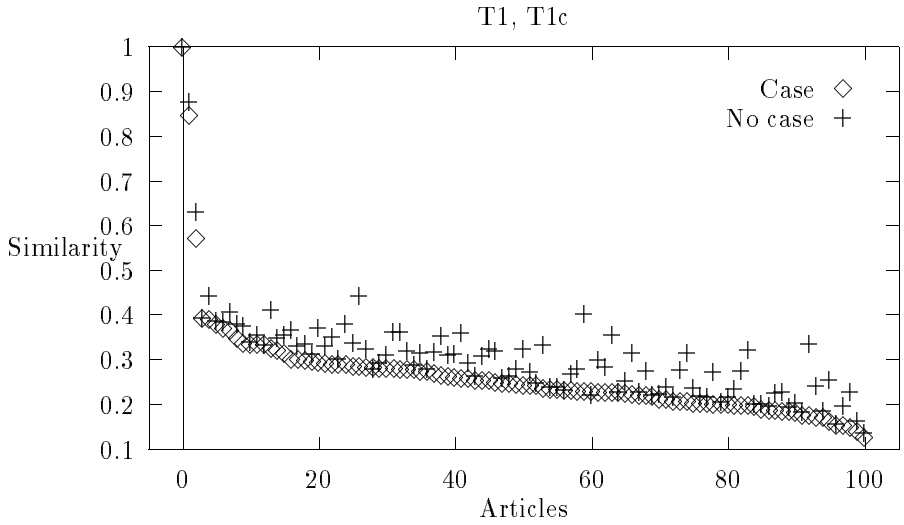

Figure 11: Cosine, one word, difference on case.

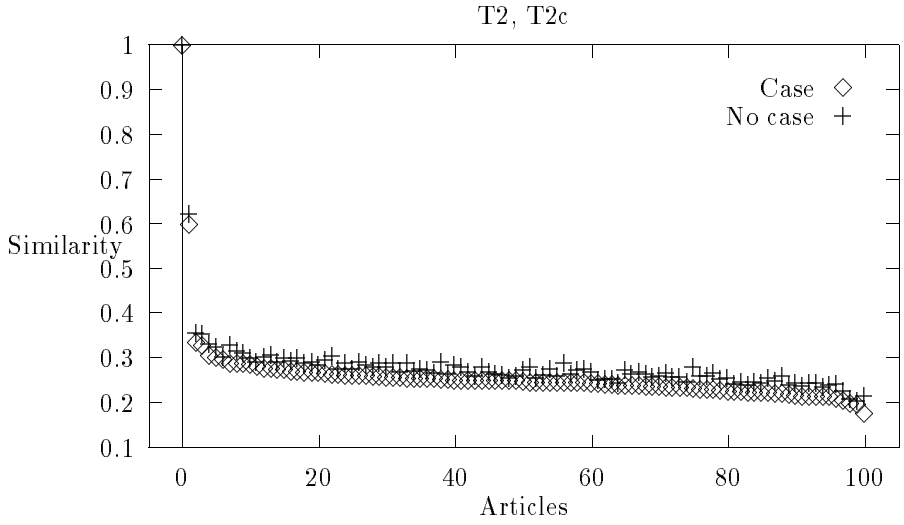

Figure 12: Sub-string, one word, difference on case.

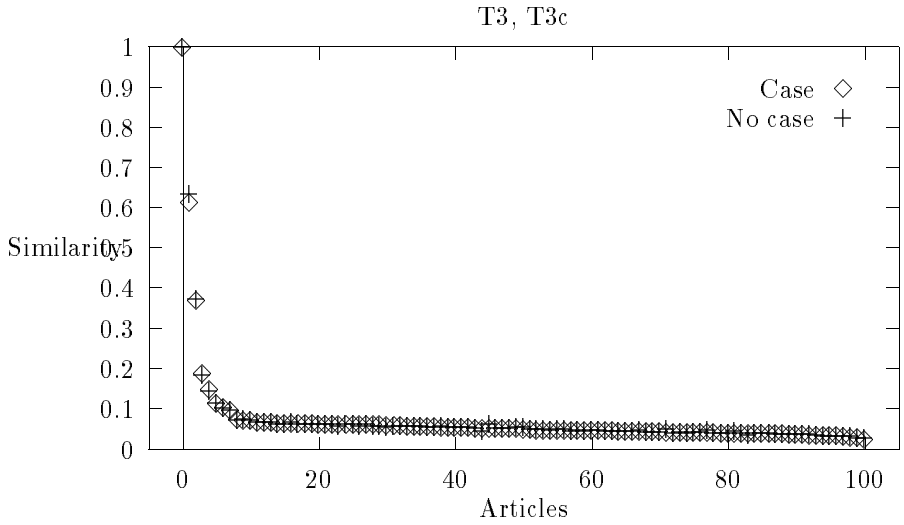

Figure 13: Cosine, two word phrase, difference on case.

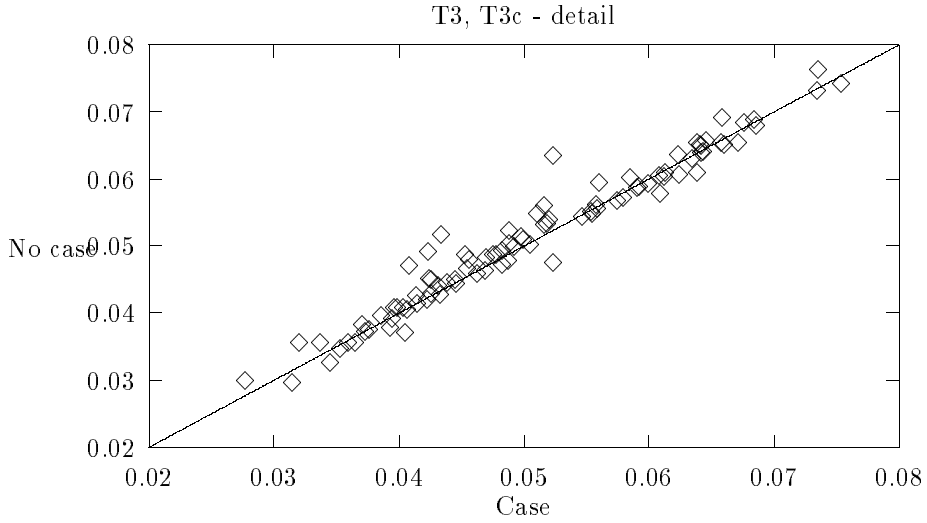

Figure 14: Cosine, two word phrase, difference on case, detail.

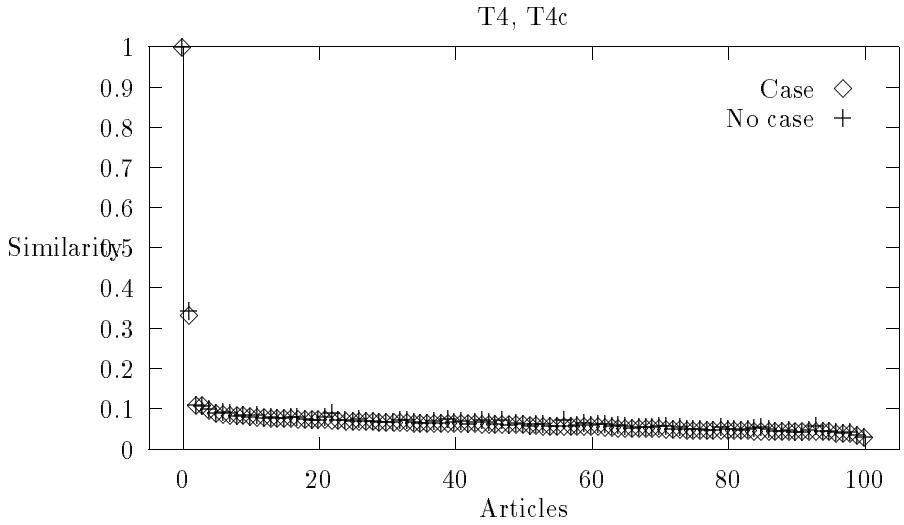

Figure 15: Sub-string, two word phrase, difference on case.

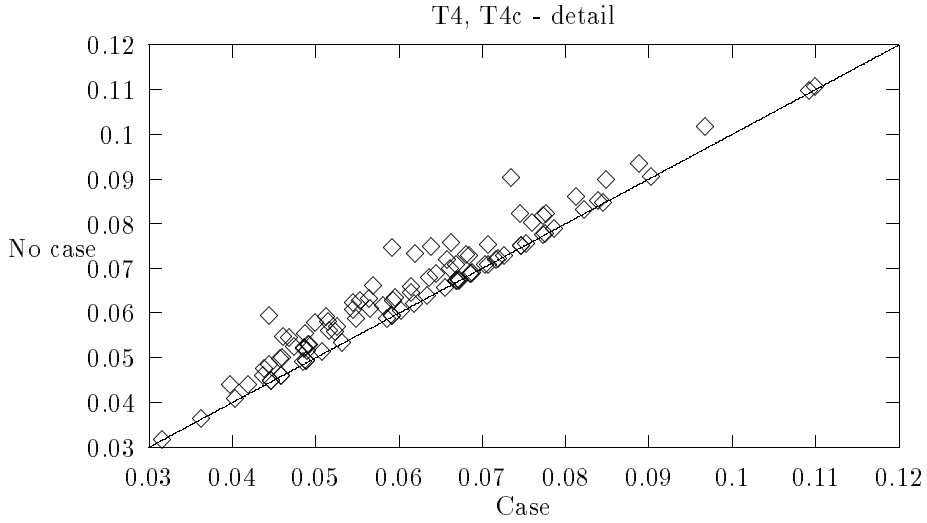

Figure 16: Sub-string, two word phrase, difference on case, detail.

References

[MS94] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 272–281. Springer-Verlag, 1994.

[SM83] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.