Part of Speech Tagging for Text Clustering in Swedish

Magnus Rosell
KTH CSC
Stockholm, Sweden
rosell@csc.kth.se

Abstract

Text clustering could be very useful both as an intermediate step in a large natural language processing system and as a tool in its own right. The result of a clustering algorithm is dependent on the text representation that is used. Swedish has a fairly rich morphology and a large number of homographs. This possibly leads to problems in Information Retrieval in general. We investigate the impact on text clustering of adding the part-of-speech-tag to all words in the the common term-by-document matrix.

The experiments are carried out on a few different text sets. None of them give any evidence that part-of-speech tags improve results. However, to represent texts using only nouns and proper names gives a smaller representation without worsen results. In a few experiments this smaller representation gives better results.

We also investigate the effect of lemmatization and the use of a stoplist, both of which improves results significantly in some cases.

1 Introduction

Text clustering (see for instance Manning et al. (2008)) aims at dividing a set of texts into groups with coherent content without knowledge of any predefined categories. The result of a clustering could be useful in many different circumstances: it can be used as an intermediate step in a bigger system, or as a tool in its own right, to facilitate exploration of search engine results (Zamir et al., 1997) or for any text set (Cutting et al., 1992).

The result of clustering algorithms is dependent on a definition of a (dis)similarity between the objects. For text clustering the similarity is usually defined via a representation of the texts using some or all the words/tokens that appear in them. Two texts are typically defined as similar if they use the same words. Which words/tokens that are used and how they are preprocessed can have a great effect on the result.

Lemmatization or stemming allows us to treat several related tokens as the same, leading to an increased similarity between texts, using the different forms of a word. Part-of-speech (PoS) tagging can be used to achieve the opposite; separate homographs so that texts are not defined similar when they are using the different meanings of a token.

The rest of this paper is organized as follows. Sections 2 through 4 gives a background to the experiments that we have conducted and present in Section 5. Finally, in Section 6 we summarize and draw some conclusions.

2 Information Retrieval

In Information Retrieval (IR) texts are represented in the common vector space model, see any introductory text, for instance (Manning et al., 2008). Each element of a term-document-matrix is assigned a weight, modeling the importance of the corresponding term to the document. There are several weighting schemes; we use a tf*idf weighting scheme. The similarity between texts (in a search engine: a query and a text) is modeled by a measure that compare their corresponding columns in the matrix. We use the common cosine measure, the cosine of the angle between the vectors.

When building the representation a few preprocessing steps are usually applied after tokenization, depending on the application. Common terms are included in a stoplist and removed, as these usually not contribute to the similarity calculations, being present in many texts. Modern search engines do not use them at all since the

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 150–157

original motivation was to save storage space.

Token (or term) normalization, further, reduces classes of related terms to common representatives to increase similarity between texts that contain these. This includes a predetermined way to handle such things as capital letters, hyphenations, abbreviations, etc. From a linguistic point of view, the most interesting part of term normalization is the use of stemming or lemmatization to collapse morphological variants of a word. Stemming is a more ad hoc method that removes affixes and may reduce word derivations having different parts of speech into the same so called stem, while lemmatization refers to replacing each token with its proper lemma. The effect of using stemming on English texts for search engines is somewhat debated. Some studies have shown improvements, while others even a decrease in performance. There have been improvements reported when using stemming and/or lemmatization for several other languages.

In 2001 Hedlund et al. observed that Swedish was poorly known from an IR perspective. They identify a few properties of the Swedish language that are potential problems (as compared to for instance English):

1. The rather rich morphology (inflectional and derivative).

2. The frequent formation of compounds, which appear as one token. (Of words remaining after the use of a stoplist 10 % are compounds, meaning that more than 20 % of the interesting morphemes are found in compounds.)

3. The high frequency of homographic words. (65% of words in running text)

To address these problems they suggest using natural language processing (NLP) tools: word normalization (stemming or lemmatization) for the morphological variation, compound splitting to extract the information in the parts, and part-of-speech tagging with gender for nouns to disambiguate homographic words. However, search queries are usually short and can be hard to part-of-speech tag correctly.

An IR system for Swedish has to take these issues into consideration. There has been a lot of work done on search engines for both mono and cross language retrieval in recent years. A big comparative study of several European languages is (Hollink et al., 2004). We feel a bit sceptic about the results for Swedish (and Finnish) since they report a substantial increase in performance when removing diacritic characters, indicating that the system does not handle the language very well. They also report substantial improvements using stemming and compound splitting for Swedish.

There are also a lot of studies in CLEF[1] (The Cross-Language Evaluation Forum) that include Swedish, several of which report improvements using morphological analysis.

Carlberger et el. (2001) saw an increase in search engine precision and recall on a newspaper text set when using stemming as compared to not using it. Ahlgren and Kekäläinen (2007) study several user scenarios on newspaper texts and report improvements for morphological analysis, word truncation, and compound splitting.

The results for search engines do not necessarily hold true for other IR methods, such as text clustering.

3 Text Clustering

The vector space model described in Section 2 can be used for text clustering. The reason for doing this is to define similarity between texts and/or groups of texts. Therefore it is not necessary to keep all tokens as in a search engine, where the goal is to be able to retrieve texts containing certain tokens. Hence, the results for search engines are not necessarily valid for text clustering.

Text clustering of Swedish texts has been investigated with respect to stemming and compound splitting (Rosell, 2003) and the use of nominal phrases in the representation (Rosell and Velupillai, 2005). Stemming seems to improve results, but the improvement is small. Compound splitting improves results, but the use of nominal phrases in the representation does not.

We use the K-Means clustering algorithm, see (Jain et al., 1999) for instance. It is fast and efficient and iteratively improves on k centroids (mean vectors) that represent k clusters. In each iteration each text is assigned to the group with the most similar centroid[2]. The algorithm stops when no text changes cluster between iterations. In the experiments presented here we stop after 20 itera-

[1] http://clef-campaign.org/

[2] We do not normalize the centroids when calculating similarity, leading to the average similarity between the text and all texts in the cluster.

tions, as the early iterations contribute more to the result.

In K-Means clustering each centroid contains all terms appearing in all texts of its cluster: terms with high weight in a centroid co-occur a lot in the cluster. If there is coherent content groups in the text set K-Means can find them or something related to them via centroids of cooccurring terms.

Homographs with several meanings may appear in several centroids and be disambiguated by the other terms. Synonyms will likely co-occur with the same words, and hence be present in the same centroid(s). In this work we investigate if these effects can be improved by separating homographs of different parts-of-speech.

4 Clustering Evaluation

Evaluation of text clustering can be either internal or external. Internal measures defines the quality of a clustering using the same information available to the clustering algorithm; the representation and/or similarity measure. As we evaluate different representations these are not appropriate here.

External evaluation can be performed by studying the effect of a clustering on a system that uses clustering as an intermediate step, by asking users for their opinions on the clustering result, or by comparing the result to a known categorization. The later is the easiest, fastest, and least expensive.

Among external measures based on comparisons of a clustering C with a known categorization K the mutual information (MI) is good since it compares the entire distribution of texts over the clusters to the entire distribution of texts over the categories (Strehl, 2002):

$$MI(C,K) = \sum_{i=1}^{\gamma}\sum_{j=1}^{\kappa}\frac{m_i^{(j)}}{n}\log(\frac{m_i^{(j)}n}{n_i n^{(j)}}),$$

where γ and κ are the numbers of clusters and categories, n the total number of texts, n_i the number of texts in cluster $c_i \in C$, $n^{(j)}$ the number of texts in category $k^{(j)} \in K$, and $m_i^{(j)}$ the number of texts in both cluster c_i and category $k^{(j)}$.

The normalized mutual information (NMI) takes the distributions of the texts over the clustering and the categorization into account (Strehl and Ghosh, 2003):

$$NMI(C,K) = \frac{2MI(C,K)}{\sqrt{H(C)H(K)}},$$

where $H(C) = -\sum_{i=1}^{\gamma}\frac{n_i}{n}\log\frac{n_i}{n}$ is the entropy for the distribution of texts over the clusters, and $H(K)$ similarly. This makes comparison of evaluations of different clusterings compared to different categorizations theoretically possible. However, the mutual information can never take the inherent linguistic structure of different text sets into account; although comparable in both size of the entire set and distribution over categories, two text sets need not be similarly hard to cluster!

5 Experiments

We have clustered several text sets, see Section 5.1, with several different text representations described in Section 5.2 to a few different numbers of clusters (5, 10, 50) using the K-Means algorithm. All results presented here are average results over 20 runs with standard deviations in parenthesis.

5.1 Text Sets

We have used the following text sets:

KTH News Corpus (Hassel, 2001) is a set of downloaded news texts. The news are from different sources, some of which have a categorization. For the newspapers *Aftonbladet* and *Dagens Nyheter* the texts are categorized into five sections: Domestic/Sweden, Foreign/World, Economy, Culture/Entertainment, and Sports. We have extracted some small text sets from these:

A is some of the texts with 20 or more words from Aftonbladet.

DN is all of the texts with 20 or more words from Dagens Nyheter.

Occ comes from a questionnaire in The Swedish Twin Registry[3]. This text set is the free text answers from 1998 and 2002 to a question about occupation given to the twins born in and before 1958. All answers were categorized by Statistics Sweden[4] (SCB) according to two hierarchical occupation classification systems:

[3] The largest twin registry in the world, containing information about more than 140 000 twins. See (Lichtenstein et al., 2002; Lichtenstein et al., 2006) for a description of the contents and some findings that have come from it and http://www.meb.ki.se/twinreg/index_en.html for more information.

[4] http://www.scb.se

AMSYK is used by AMS (The Swedish National Labour Market Administration) and is based on ISCO88 (The International Standard Classification of Occupations).

YK80 was used in The Swedish Population and Housing Census 1980.

Table 1 gives the number of categories on each of the levels in the classification systems. For the evaluation of these experiments we have used the second level of both.

	L1	L2	L3	L4	L5
AMSYK	11	28	114	361	969
YK80	12	59	288		

Table 1: The Occupation Classification Systems (number of categories per level)

	Text Sets		
	A	DN	Occ
Texts	2424	6395	41949
Categories	5	5	28, 59
$H(K)/\log(\kappa)$	1.00	0.97	0.90, 0.83
Word Forms	12071	37725	17594
Forms/Text	52.29	97.41	15.60
Texts/Form	10.50	16.51	37.20
Lemmas	9050	26451	13873
Lemmas/Text	48.84	88.13	13.70
Texts/Lemma	13.08	21.31	41.29

Table 2: Text Set Statistics

We have used the grammar checking program `Granska`[5] (Domeij et al., 1999) for tokenization, lemmatization, and to tag each word with its part-of-speech. Table 2 gives some statistics for the text sets after preprocessing to word forms (including delimiters) and lemmas. The number of texts, tokens, and the average number of unique token per text and texts per unique token. We also give the number of categories and the "evenness" of the categorization: $H(K)/\log(\kappa)$, which is 1 for a categorization where all categories have equal size, and lower for other cases.

As can be seen there is a significant decrease in tokens when using lemmas instead of word forms. Even if this does not improve the results it improves the storage requirements for the representations.

5.2 Representation

We have evaluated several different representations, which we describe briefly here. In the next section (Section 5.3) we present the results.

`Granska` outputs among other things a tokenization that contains word forms, lemmas, the part-of-speech for each token, and some delimiters. The part-of-speech classes are given in Table 3 and is an adaption (Carlberger and Kann, 1999) of the the tag set in the Stockholm-Umeå Corpus (SUC) (Källgren and Eriksson, 1993).

We have used all the tokens in the representation we call *Full* with either word forms or lemmas (*Word Form* and *Lemma* in the tables). To reduce the Full representation one can use either a stoplist or only consider tokens that get a proper wordclass as their part of speech. The *All wordclasses* representation uses all tokens with these, except the delimiters.

For the *Stoplist* representation we removed tokens according to the Swedish stoplist of the snowball stemmer[6], plus all numbers, and words shorter than three characters and longer than 20.

To separate homographs by their part-of-speech we create new features by concatenating the lemma with its part-of-speech tag (*Lemma* + *PoS*), for instance: "och_kn", "spela_vb", "mittback_nn". We compare the results for this representation to the one using only the lemma. To separate even more homographs we use the gender for nouns as well (*Lemma* + *PoS* + *Gender*).

Most parts of speech in Table 3 contain only words that are usually in a stoplist. We have concentrated on the largest wordclasses, as these are also the ones that convey content in an obvious way. In the result tables we indicate which we have used by the abbreviations in Table 3.

When the representation is constructed we remove terms that appear in only one text as these do not contribute to the similarity calculations. We also remove texts that only contain one term.

5.3 Results

We present some results for text set DN in Table 4, and some of the results for text set Occ evaluated against the second level of the AMSYK categorization system in Table 5. The results for text set

[5] http://www.nada.kth.se/theory/projects/granska/

[6] http://snowball.tartarus.org/

Abbreviation	Part-of-Speech	Example
nn	noun	bil
pm	proper name	Lars
jj	adjective	grön
rg	number	12
ro	cardinal number	första
vb	verb	springa
ab	adverb	mycket
in	interjection	ja
ha	interrogative/relative adverb	när
dt	determiner	den
hd	interrogative/relative determiner	vilken
ps	possesive pronoun	hennes
hs	interrogative/relative possessive	vems
pn	pronoun	hon
hp	interrogative/relative pronoun	vem
sn	subordinating conjunction	om
kn	coordinating conjunction	och
pp	preposition	till
pc	participle	springande
pl	particle	om
uo	foreign word	the
an	abbreviation	d.v.s.
ie	verb base form marker	att
dl	delimiter	.

Table 3: Part-of-Speech Tags used in `Granska`

A are very similar to the ones for DN, and also the results on text set Occ evaluated against the YK80 categorization system (level 2) are very similar to the results evaluated to the AMSYK categorization.

The tables are divided into sections vertically for different numbers of clusters and horizontally for which features are used in the text representation: Word Form, Lemma, Lemma + PoS, or Lemma + PoS + Gender. Other aspects of the representation are presented as rows; which of the features are used in the representation.

The result of each experiment (20 K-Means clusterings of a particular representation) is presented with two values: the average NMI with standard deviations in parenthesis, and the number of features the representation gives rise to. As we remove texts that have one or fewer features some of the clustering are performed on fewer texts than are presented in Table 2. The number of texts that are removed are under on per cent in all cases.

Most differences are well within the standard deviations and should therefore not be considered significant. The representations are kept constant in the experiments; the varying results are due to the indeterministic K-Means algorithm.

5.4 Discussion

Our attempt to enhance the representation by introducing the part-of-speech tags (and gender) fails miserably. There are no interesting tendencies pointing to any improvements compared to using only lemmas, see Tables 4:b, 4:c, and 5:b. The effect of keeping only some parts-of-speech in the representation is not surprising: adjectives, verbs, and adverbs are not very good, while the nouns and proper names are as good on their own as all parts-of-speeches together. For five clusters on the Occ text set it is even better to only keep the large word classes than using them all (Table 5:b).

We have not tried a combination of the word form and the part-of-speech tag. This would have resulted in a representation with even more features, but might have given better results than the word forms on their own.

The lemmatization might address the homograph problem to some extent in addition to the morphological variants. An other explanation is that the cooccurence statistics gathered in the centroids is quite effective in separating homographs, and is not very dependent on which representation is used. Regardless of whether any of these two explanations are true, a representation extended with PoS tags does not improve results.

The comparison between word form and lemma representation in Tables 4:a and 5:a contains some interesting results. It is almost always beneficial to use lemmatization, and most times it improves results a lot. For text set DN it does not improve results significantly when clustering to five clus-

		Word Form		Lemma	
Clusters	Representation	NMI	Features	NMI	Features
5	Full	0.44 (0.05)	37725	0.52 (0.05)	26466
	Stoplist	0.52 (0.04)	35888	0.51 (0.04)	25604
	All wordclasses	0.47 (0.06)	37705	0.49 (0.04)	26451
50	Full	0.28 (0.01)	37725	0.35 (0.01)	26466
	Stoplist	0.28 (0.01)	35888	0.35 (0.01)	25604
	All wordclasses	0.28 (0.01)	37705	0.35 (0.01)	26451

a) Word Form vs. Lemma

		Lemma		Lemma + PoS	
Clusters	Representation	NMI	Features	NMI	Features
5	All wordclasses	0.49 (0.04)	26451	0.51 (0.04)	27532
	nn, pm, jj, vb, ab	0.50 (0.04)	25923	0.52 (0.05)	26767
	nn, pm	0.54 (0.04)	19507	0.55 (0.05)	19940
	jj, vb, ab	0.28 (0.02)	6729	0.29 (0.02)	6827
	jj, ab	0.20 (0.01)	4231	0.19 (0.01)	4285
	vb	0.27 (0.02)	2542	0.27 (0.02)	2542
50	All wordclasses	0.35 (0.01)	26451	0.34 (0.01)	27532
	nn, pm, jj, vb, ab	0.35 (0.01)	25923	0.34 (0.01)	26767
	nn, pm	0.37 (0.01)	19507	0.37 (0.01)	19940
	jj, vb, ab	0.24 (0.01)	6729	0.24 (0.01)	6827
	jj, ab	0.17 (0.01)	4231	0.17 (0.00)	4285
	vb	0.19 (0.00)	2542	0.19 (0.01)	2542

b) Lemma vs. Lemma + PoS

		Lemma + PoS + Gender	
Clusters	Representation	NMI	Features
5	All wordclasses	0.52 (0.04)	27612
	nn, pm, jj, vb, ab	0.50 (0.05)	26847
	nn, pm	0.51 (0.06)	20020
50	All wordclasses	0.34 (0.01)	27612
	nn, pm, jj, vb, ab	0.35 (0.01)	26847
	nn, pm	0.37 (0.01)	20020

c) Lemma + PoS + Gender

Table 4: Some Results for Text Set DN (about 6400 news articles)

ters, but it does not worsen results. The biggest improvement is for text set Occ clustered to five clusters, more than 50 % on average (standard deviation of about 20 %).

The stoplist improves results for text set Occ, but not for DN. It is particularly in combination with lemmatization, when clustering to few clusters that this can be seen. Perhaps the stop words obscure the representation more in the short texts of Occ. To use only the tokens that have proper wordclasses (All wordclasses) does not improve results. The Full representation does, however, not contain many other tokens in the first place.

Lemmatization effects all words/tokens in the representation. We expected that this global influence should be more obvious in results than the use of a stoplist, which is more local. However, the stop words adds noise; making all texts a bit similar, something which seems to be more important for short texts.

The clustering achieves better results when the number of clusters are roughly the same as the number of categories in the categorization used for the evaluation[7], regardless of the representation. It seems hard to improve results for this “optimal” number of clusters using the different representations we try here.

In these experiments we have used almost all words/tokens as features. It is possible to remove a lot of the features without getting worse results. We have tried a few versions were we remove words that appear in few documents. The general tendencies are still the same. Most notably there is nothing to be gained from using the part of speech tags.

Although results do not always improve with the use of lemmatization and a stoplist they never

[7] This is not surprising, considering the definition of NMI. For measures considering only the quality of any single cluster (not the entire clustering) the quality usually improves with more and smaller clusters.

		Word Form		Lemma	
Clusters	Representation	NMI	Features	NMI	Features
5	Full	0.10 (0.02)	17594	0.15 (0.02)	13916
	Stoplist	0.13 (0.02)	16378	0.25 (0.02)	13200
	All wordclasses	0.09 (0.01)	17546	0.15 (0.02)	13873
50	Full	0.25 (0.01)	17594	0.29 (0.01)	13916
	Stoplist	0.29 (0.01)	16378	0.33 (0.01)	13200
	All wordclasses	0.26 (0.01)	17546	0.30 (0.01)	13873

a) Word Form vs. Lemma

		Lemma		Lemma + PoS	
Clusters	Representation	NMI	Features	NMI	Features
5	All wordclasses	0.15 (0.02)	13873	0.15 (0.02)	14151
	nn, pm, jj, vb, ab	0.20 (0.02)	13565	0.20 (0.03)	13704
	nn, pm	0.23 (0.02)	10834	0.23 (0.01)	10841
50	All wordclasses	0.30 (0.01)	13873	0.30 (0.01)	14151
	nn, pm, jj, vb, ab	0.31 (0.01)	13565	0.31 (0.01)	13704
	nn, pm	0.31 (0.01)	10834	0.31 (0.01)	10841

b) Lemma vs. Lemma + PoS

Table 5: Some Results for Text Set Occ (about 42000 short texts)

deteriorate. On the other hand sometimes results improve a great deal. If a minimal representation is required one should consider using only nouns and proper names.

6 Conclusions and Further Work

We conclude that part of speech tagging does not improve results for text clustering of Swedish texts. However, to use only nouns and proper names in the representation often leads to results comparable to using all words, and may decrease the number of features significantly.

Lemmatization improves results a lot in several experiments. To use a stoplist improves results sometimes; in our experiments for short texts.

The cooccurence information in the K-Means centroids is obviously very good at handling homographs as no improvement in clustering results was achieved when introducing lemma-PoS-tag features.

As nouns seems to be very important for clustering, pronoun resolution could perhaps be interesting. However, it would just alter the weighting for the nouns and thus not affect the similarity between texts quite as radically as lemmatization and part of speech tagging.

References

P. Ahlgren and J. Kekäläinen. 2007. Indexing strategies for swedish full text retrieval under different user scenarios. *Inf. Process. Manage.*, 43(1):81–102.

J. Carlberger and V. Kann. 1999. Implementing an efficient part-of-speech tagger. *Softw. Pract. Exper.*, 29(9):815–832.

J. Carlberger, H. Dalianis, M. Hassel, and O. Knutsson. 2001. Improving precision in information retrieval for Swedish using stemming. In *Proc. 13th Nordic Conf. on Comp. Ling. – NODALIDA '01*.

D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. 15th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*.

R. Domeij, O. Knutsson, J. Carlberger, and V. Kann. 1999. Granska – an efficient hybrid system for Swedish grammar checking. In *Proc. 12th Nordic Conf. on Comp. Ling. – NODALIDA '99*.

M. Hassel. 2001. Automatic construction of a Swedish news corpus. In *Proc. 13th Nordic Conf. on Comp. Ling. – NODALIDA '01*.

T. Hedlund, A. Pirkola, and K. Järvelin. 2001. Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. *Information Processing & Management*, 37(1):147–161.

V. Hollink, J. Kamps, C. Monz, and M. De Rijke. 2004. Monolingual document retrieval for european languages. *Inf. Retr.*, 7(1-2):33–52.

A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

G. Källgren and G. Eriksson. 1993. The linguistic annotation system of the stockholm: Umeå corpus project. In *Proceedings of the sixth conference on*

European chapter of the Association for Computational Linguistics, pages 470–470, Morristown, NJ, USA. Association for Computational Linguistics.

P. Lichtenstein, U. De faire, B. Floderus, M. Svartengren, P. Svedberg, and N. L. Pedersen. 2002. The Swedish twin registry: a unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine*, 252:184–205.

P. Lichtenstein, P. F. Sullivan, S. Cnattingius, M. Gatz, S. Johansson, E. Carlstrom, C. Bjork, M. Svartengren, A. Wolk, L. Klareskog, U. de Faire M. Schalling, J. Palmgren, and N. L. Pedersen. 2006. The Swedish twin registry in the third millennium: An update. *Twin Research and Human Genetics*, 9(6):875–882.

C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

M. Rosell and S. Velupillai. 2005. The impact of phrases in document clustering for Swedish. In *Proc. 15th Nordic Conf. on Comp. Ling. – NODALIDA '05*.

M. Rosell. 2003. Improving clustering of Swedish newspaper articles using stemming and compound splitting. In *Proc. 14th Nordic Conf. on Comp. Ling. – NODALIDA '03*.

A. Strehl and J. Ghosh. 2003. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617.

A. Strehl. 2002. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. Ph.D. thesis, The University of Texas at Austin.

O. Zamir, O. Etzioni, O. Madani, and R. M. Karp. 1997. Fast and intuitive clustering of web documents. In *Knowledge Discovery and Data Mining*, pages 287–290.