Matching Similarity for Keyword-Based Clustering

Mohammad Rezaei and Pasi Fränti

University of Eastern Finland
{rezaei,franti}@cs.uef.fi

Abstract. Semantic clustering of objects such as documents, web sites and movies based on their keywords is a challenging problem. This requires a similarity measure between two sets of keywords. We present a new measure based on matching the words of two groups assuming that a similarity measure between two individual words is available. The proposed matching similarity measure avoids the problems of traditional measures including minimum, maximum and average similarities. We demonstrate that it provides better clustering than other measures in a location-based service application.

Keywords: clustering, keyword, semantic, hierarchical.

1 Introduction

Clustering has been extensively studied for text mining. Applications include customer segmentation, classification, collaborative filtering, visualization, document organization and indexing. Traditional clustering methods consider numerical and categorical data [1], but recent approaches consider also different text objects such as documents, short texts (e.g. topics and queries), phrases and terms.

Keyword-based clustering aims at grouping objects that are described by a set of *keywords* or *tags*. These include movies, services, web sites and text documents in general. We assume here that the only information available about each data object is its keywords. The keywords can be assigned manually or extracted automatically. Fig. 1 shows an example of services in a location-based application where the objects are defined by a set of keywords. For presenting an overview of available services to a user in a given area, clustering is needed.

Several methods have been proposed for the problem [2, 3, 4, 5] mostly by agglomerative clustering based on single, compete or average links. The problem is closely related to *word clustering* [6, 7, 8] but instead of single words, we have a set of words to be clustered. Both problems are based on measuring similarity between words as the basic component.

To solve clustering, we need to define a similarity (or distance) between the objects. In agglomerative methods such as *single link* and *complete link*, similarity between individual objects is sufficient, but in partitional clustering such as *k-means* and *k-medoids* cluster representative is also required to measure object-to-cluster similarity. Using semantic content, however, defining the representative of a cluster is not trivial. Fortunately, it is still possible to apply partitional clustering even without the representatives. For example, an object can be assigned to such cluster that minimizes

P. Fränti et al. (Eds.): S+SSPR 2014, LNCS 8621, pp. 193–202, 2014.

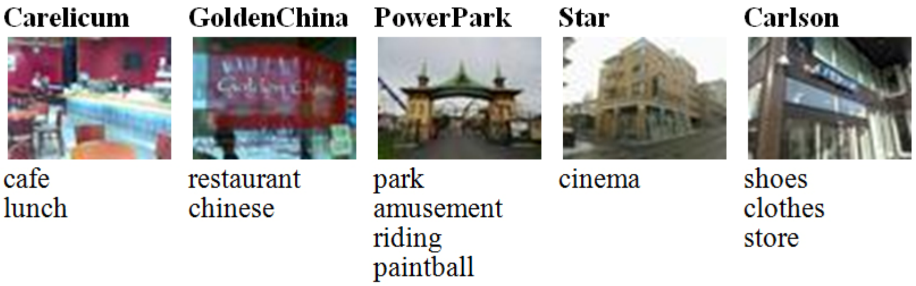

Fig. 1. Five examples of location-based services in Mopsi (http://www.uef.fi/mopsi): name of the service, representative image, and the keywords describing the service

(or maximizes) the cost function where only the similarities between objects are needed.

In this paper, we present a novel similarity measure between two sets of words, called *matching similarity*. We apply it to keyword-based clustering of services in a location-based application. Assuming that we have a measure for comparing semantic similarity between two words, the problem is to find a good measure to compare the sets of words. The proposed matching similarity solves the problem as follows. It iteratively pairs two most similar words between the objects and then repeats the process for the rest of the objects until one of the objects runs out of words. The remaining words are then matched just to their most similar counterpart in the other object.

The rest of the paper is organized as follows. In Section 2, we review existing methods for comparing the similarity of two words, and select the most suitable for our need. The new similarity measure is then introduced in Section 2. It is applied to agglomerative clustering in Section 3 with real data and compared against existing similarity measures in this context.

2 Semantic Similarity between Word Groups

In this section, we first review the existing methods for measuring semantic similarity between individual words, because it is the basic requirement for comparing two sets of words. We then study how they can be used for comparing two set of words, present the new measure called *matching similarity*, and demonstrate how it is applied in clustering of services in a location based application.

2.1 Similarity of Words

Measures for semantic similarity of words can be categorized to *corpus-based*, *search engine-based, knowledge-based* and *hybrid*. Corpus-based measures such as *point-wise mutual information* (PMI) [9] and *latent semantic analysis* (LSA) [9] define the similarity based on large corpora and term co-occurrence. Search engine-based measures such as *Google distance* are based on web counts and snippets from results of a search engine [8], [10, 11]. *Flickr distance* first searches two target words separately through the image tags and then uses image contents to calculate the distance between the two words [12].

Knowledge-based measures use lexical databases such as *WordNet* [13] and *CYC* [13], which can be considered as computational format of large amounts of human knowledge. The knowledge extraction process is very time consuming and the database depends on human judgment and it does not scale easily to new words, fields and languages [14, 15].

WordNet is a taxonomy that requires a procedure to derive the similarity score between words. Despite its limitations it has been successively used for clustering [16]. Fig. 2 illustrates a small part of WordNet hierarchy where mammal is the *least common subsumer* of wolf and hunting dog. *Depth* of a word is the number of links between it and the root word in WordNet. As an example, Wu and Palmer measure [17, 18] is defined as follows:

$$S(w_1, w_2) = \frac{2 \times depth(LCS(w_1, w_2))}{depth(w_1) + depth(w_2)} \tag{1}$$

where *LCS* is the least common subsumer of the words w_1 and w_2.

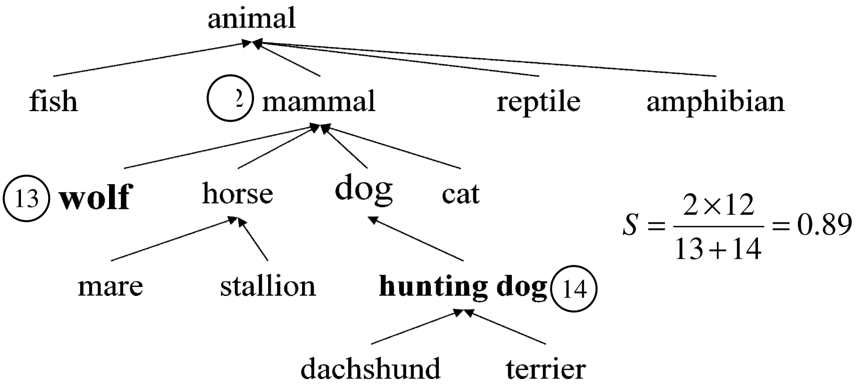

Fig. 2. Part of WordNet taxonomy; the numbers in the circles represent the depths

Jiang-Contrath [13] is a hybrid of corpus-based and knowledge-based as it extracts the information content of two words and their LCS in a corpus. Methods based on Wikipedia or similar websites are also hybrid in the sense that they use organized corpora with links between documents [19]. In the rest of the paper, we use Wu & Palmer measure due to its simplicity and reasonable results in earlier work [16].

2.2 Similarity of Word Groups

Given a measure for comparing two words, our task is to measure similarity between two sets of words. Existing measures calculate either minimum, maximum or average similarities. Minimum and maximum measures find the pair of words (one from each object) that are least (minimum) and most (maximum) similar. Average similarity considers all pairs of words and calculates their average value. Example is shown in Fig. 3, where the values are min=0.21, max=0.84, average=0.57.

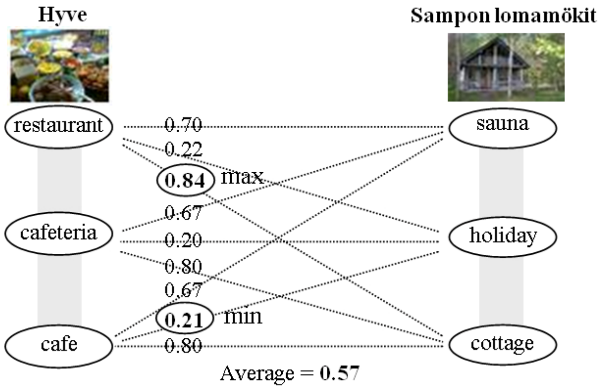

Fig. 3. Minimum and maximum similarities between two location-based services is derived by considering two keywords with minimum and maximum similarities

Now consider two objects with exactly the same keywords (100% similar) as follows:

(a) Café, lunch
(b) Café, lunch

The word similarity between Café and lunch is 0.32. The corresponding minimum, average and maximum similarity measures would result in 0.32, 0.66 and 1.00. It is therefore likely that minimum and average measures would cluster these in different groups and only maximum similarity would cluster them correctly in the same group.

Now consider the following two objects that have a common word:

(a) Book, store
(b) Cloth, store

The maximum similarity measure gives 1.00 and therefore as soon as the agglomerative algorithm processes to these objects, it clusters them in one group. However, if data contains lots of stores, they might have to be clustered differently.

The following example reveals another disadvantage of minimum similarity. These two objects should have a high similarity as their only difference is the drive-in possibility of the first service.

(a) Restaurant, lunch, pizza, kebab, café, drive-in
(b) Restaurant, lunch, pizza, kebab, café

Minimum similarity would result to *S*(drive-in, pizza)=0.03, and therefore, place the two services in different clusters.

2.3 Matching Similarity

The proposed *matching similarity* measure is based on a greedy pairing algorithm, which first finds two most similar words across the sets, and then iteratively matches next similar words. Finally, the remaining non-paired keywords (of the object with more keywords) are just matched with the most similar words in the other object. Fig. 4 illustrates the matching process between two sample objects.

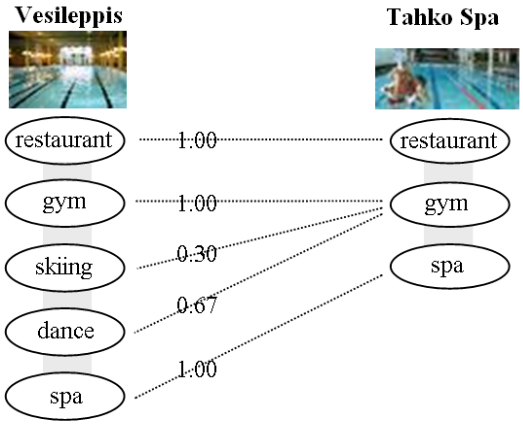

Fig. 4. Matching between the words of two objects

Consider two objects with N_1 and N_2 keywords so that $N_1 > N_2$. We define the normalized similarity between the two objects as follows:

$$S(O_1, O_2) = \frac{\sum_{i=1}^{N_1} SW(w_i^{O_1}, w_{p(i)}^{O_2})}{N_1} \tag{2}$$

where SW measures the similarity between two words, and $p(i)$ provides the index of the matched word for w_i in the other object.

The proposed measure provides more intuitive results than existing measures, and eliminates some of their disadvantages. As a straightforward property it gives the similarity 1.00 for the case of objects with same set of keywords.

3 Experiments

We study the method with Mopsi data (http://www.uef.fi/mopsi), which includes various location-tagged data such as services, photos and routes. Each service includes a set of keywords to describe what it has to offer. Both English and Finnish languages keywords have been casually used. For simplicity, we translated all Finnish words into English by Microsoft Bing translator for these experiments. Some issues raised in translation such as stop words, Finnish word converting to multiple English words, and some strange translations due to using automatic translator. We manually refined the data to remove the problematic words and the stop words.

In total, 378 services were used for evaluating the proposed measure and compare it against the following existing measures: *minimum*, *maximum* and *average similarity*. We apply complete and average link clustering algorithms as they have been widely used in different applications. Each of the clustering algorithms is performed based on three similarity measures. Here we fixed the number of clusters to 5 since our goal of clustering is to present user the main categories of services, with easy navigation to find the desired target without going through a long list. We find the natural number of clusters using *SC* criteria introduced in [16] by finding minimum *SC* value among clusterings with different number of clusters. We then display four largest clusters and put all the rest in the fifth cluster. The data and the corresponding clustering results can be found here (http://cs.uef.fi/paikka/rezaei/keywords/).

The three similarity measures of five selected services in Table 1 are demonstrated in Table 2. The first three and the last two services should be in two different clusters according to their similarities. However, both minimum and average similarities show small differences when they compare *Parturi-kampaamo Nona* with *Parturi-kampaamo Koivunoro* and *Kahvila Pikantti*, whereas the proposed matching similarity can differentiate them much better. Despite that *Parturi-kampaamo Nona* and *Parturi-kampaamo Koivunoro* have exactly the same keywords, only the matching similarity provides value 1.00 indicating perfect match.

Table 1. Similarities between five services for the measures: minimum, average and matching

Mopsi service:	A1-Parturi-kampaamo Nona	A2-Parturi-kampaamo Platina	A3-Parturi-kampaamo Koivunoro	B1-Kielo	B2-Kahvila Pikantti
Keywords;	barber hair salon	barber hair salon	barber hair salon shop	cafe cafeteria coffe lunch	lunch restaurant

Table 2. Similarity between services described in Table 1

Services	A1	A2	A3	B1	B2
	Minimum similarity				
A1	-	0.42	0.42	0.30	0.30
A2	0.42	-	0.42	0.30	0.30
A3	0.42	0.42	-	0.30	0.30
B1	0.30	0.30	0.30	-	0.32
B2	0.30	0.30	0.30	0.32	-
	Average similarity				
A1	-	0.67	0.67	0.47	0.51
A2	0.67	-	0.67	0.47	0.51
A3	0.67	0.67	-	0.48	0.51
B1	0.47	0.47	0.48	-	0.63
B2	0.51	0.51	0.51	0.63	-
	Matching similarity				
A1	-	1.00	0.99	0.57	0.56
A2	1.00	-	0.99	0.57	0.56
A3	0.99	0.99	-	0.55	0.56
B1	0.57	0.57	0.55	-	0.90
B2	0.56	0.56	0.56	0.90	-

In general, the problems of minimum and average similarities are observable in the clustering results both for complete and average link. Several services with the same set of keywords (barber, hair, salon) are clustered together, and a service with the same keywords has its own cluster when complete link clustering is applied with minimum similarity measure. Average link method clusters the services with these keywords correctly but for services with other keywords (sauna, holiday, cottage), it clusters them in different groups even when using average similarity. This problem does not happen with matching similarity.

Another observation of minimum similarity with complete link clustering is that there appear many clusters with only one object, and a very large cluster that contains most of the other objects. Matching similarity leads to more balanced clusters with both algorithms. Interestingly, it also produces almost the same clusters with the two different clustering methods.

For more extensive objective testing, we should have a ground truth for the wanted clustering but this is not currently available as it is non-trivial to construct. We therefore make indirect comparison by using the *SC* criterion from [16]. The assumption here is that the smaller the value, the better is the clustering. Fig. 5 summarizes the SC-values for different number of clusters. The overall minima for complete link and average link are 131, 86, 146 (minimum, average and matching similarities) and 279, 96 and 140, respectively. Our method provides always the minimum *SC* value. The sizes of 4 biggest clusters in each case are listed in Table 3.

Table 3. The sizes of the four largest clusters for complete and average link clustering

Complete link				
Similarity:	Sizes of 4 biggest clusters			
Minimum	106	88	18	18
Average	44	22	20	19
Matching	27	23	19	17
Average link				
Similarity:	Sizes of 4 biggest clusters			
Minimum	22	12	10	8
Average	128	41	34	17
Matching	27	23	17	17

The effectiveness of the proposed method for displaying data with limited number of clusters still exists. The number of clusters is too large for practical use and we need to improve the clustering validity index to find larger clusters but without creating meaningless clusters. We also observed some issues in clustering that originate from the similarity measure of two words, which implies that better similarity measure would also be useful.

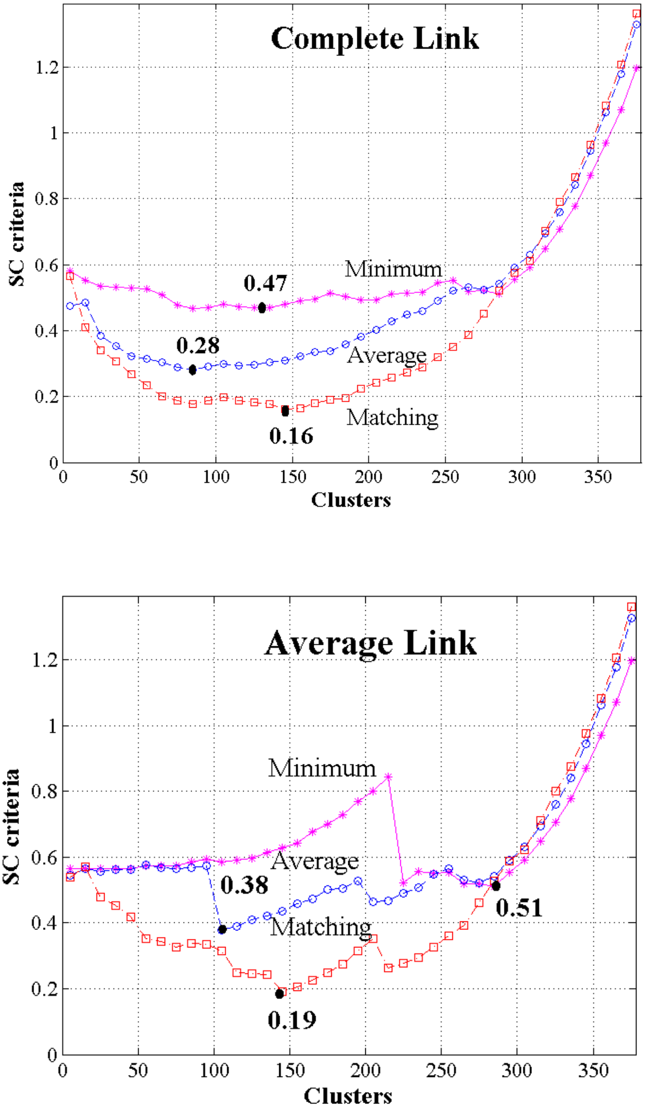

Fig. 5. Complete link and average link clustering using three similarity measures

4 Conclusion

A new measure called matching similarity was proposed for comparing two groups of words. It has simple intuitive logic and it avoids the problems of the considered minimum, maximum and average similarity measures, which fail to give proper results with rather simple cases. Comparative evaluation on a real data with SC criterion demonstrates that the method outperforms the existing methods in all cases, and by a clear marginal. A limitation of the method is that it depends on the semantic similarity measure between two words. As future work, we plan to generalize the matching similarity to other clustering algorithms such as k-means and k-medoids.

Acknowledgements. This research has been supported by MOPIS project and partially by Nokia Foundation grant.

References

1. Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In: Mining Text Data, pp. 77–128. Springer US (2012)
2. Ricca, F., Pianta, E., Tonella, P., Girardi, C.: Improving Web site understanding with keyword-based clustering. Journal of Software Maintenance and Evolution: Research and Practice 20(1), 1–29 (2008)
3. Hasan, B., Korukoglu, S.: Analysis and Clustering of Movie Genres. Journal of Computing 3(10) (2011)
4. Ricca, F., Tonella, P., Girardi, C., Pianta, E.: An empirical study on keyword-based web site clustering. In: Proceedings of the 12th IEEE International Workshop on Program Comprehension. IEEE (2004)
5. Kang, S.S.: Keyword-based document clustering. In: Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, vol. 11. Association for Computational Linguistics (2003)
6. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of English words. In: Proceedings of the 31st Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics (1993)
7. Ushioda, A., Kawasaki, J.: Hierarchical clustering of words and application to NLP tasks. In: Proceedings of the Fourth Workshop on Very Large Corpora (1996)
8. Matsuo, Y., Sakaki, T., Uchiyama, K., Ishizuka, M.: Graph-based word clustering using a web search engine. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2006)
9. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI, vol. 6 (2006)
10. Cilibrasi, R.L., Vitanyi, P.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
11. Bollegala, D., Matsuo, Y., Ishizuka, M.: A web search engine-based approach to measure semantic similarity between words. IEEE Transactions on Knowledge and Data Engineering 23(7), 977–990 (2011)
12. Wu, L., et al.: Flickr distance: a relationship measure for visual concepts. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(5), 863–875 (2012)
13. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics 32(1), 13–47 (2006)
14. Kaur, I., Hornof, A.J.: A comparison of LSA, WordNet and PMI-IR for predicting user click behavior. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM (2005)

15. Gledson, A., Keane, J.: Using web-search results to measure word-group similarity. In: Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1. Association for Computational Linguistics (2008)
16. Zhao, Q., Rezaei, M., Chen, H., Franti, P.: Keyword clustering for automatic categorization. In: 2012 21st International Conference on Pattern Recognition (ICPR). IEEE (2012)
17. Michael Pucher, F.T.W.: Performance Evaluation of WordNet-based Semantic Relatedness Measures for Word Prediction in Conversational Speech (2004)
18. Markines, B., et al.: Evaluating similarity measures for emergent semantics of social tagging. In: Proceedings of the 18th International Conference on World Wide Web. ACM (2009)
19. Berry, M.W., Dumais, S.T., O'Brien, G.W.: Short text clustering by finding core terms. Knowledge and Information Systems 27(3), 345–365 (2011)