Clustering of Short Commercial Documents for the Web

Moreno Carullo, Elisabetta Binaghi and Ignazio Gallo
Università degli Studi dell'Insubria
Dipartimento di Informatica e Comunicazione
21100 Varese, Italy
moreno.carullo@uninsubria.it

Nicola Lamberti
7pixel srl
20082 Binasco, Italy
nicola@trovaprezzi.it

Abstract

Document clustering techniques have been applied in several areas, with the web as one of the most recent and influent. Both general-purpose and text-oriented techniques exist and can be used to cluster a collection of documents in many ways. In this work we propose an online, single-pass document clustering model that can be combined with a variety of text-oriented similarity measures. An experimental evaluation of the proposed model was conducted in the e-commerce domain. Performances were measured using a clustering-oriented metric based on F-Measure and compared with those obtained by other well-known approaches.

1. Introduction

The document clustering process [3], wich is an instance of the cluster analysis paradigm, takes into account the problem of dividing a collection of documents $D = \{d_1, \dots, d_n\}$ into subsets $\{D_i, \dots, D_c\}$ such that all the $d_j \in D_i$ are more similar one to each other in regard of a given similarity measure S than with other documents outside the cluster.

Within the flat and hard clustering paradigm, the general purpose K-means algorithm [8, 5] has been successfully applied in the document analysis domain. In [15] the K-Means, put in comparison with the well-known document-oriented Hierarchical Agglomerative Clustering Method (HACM) [3], shows a competitive behavior. The flat, hard Single Pass algorithm [3] has its roots in the early works with document clustering [11] where strict time and space resource constraints played a major role.

Unsupervised neural models are deeply investigated for solving clustering problems; in particular Self-Organizing Maps [18] have been employed for document clustering tasks [7] and are particularly fitted when a meaningful and browsable 2D map of the considered document collection is required.

The representation of documents is a key aspect in all the document clustering approaches, strongly affecting the type of similarity measures used and consequently the overall performances. A widely adopted representation is the Vector Space Model (VSM) [13] usually complemented with dimensionality reduction techniques [1, 6, 17]; it treats permutations of terms like equivalent texts and lacks the positional information and mutual dependencies between terms. When this information is relevant, alternative encodings are proposed such as n-gram models [10], mixed techniques like in Indri [16], direct representation of the documents in the original symbolic form and phrase based approaches [14]; all these allow to exploit domain-specific similarity measures.

Interesting new scenarios emerged within the Web, like the collection clustering of news from the Internet seen in the Columbia NewsBlaster [9] (now Google News) or the clustering of web search results [2], feature great amount of data that has to be processed on the fly or with strict computing times. Clustering quality and speed have to be considered carefully when addressing such problems.

In this context online algorithms where the document collection is read once and very low comparisons between the elements are performed, have great value. In this work we propose and experimentally investigate a new document clustering algorithm attempting to optimize the balancing between speed and clustering quality.

The solutions adopted, particularly oriented to short documents management, are evaluated within the e-commerce domain, on commercial offers i.e. textual descriptions of product offerings on e-commerce websites.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

2. The proposed document clustering approach

In this section we describe an heuristics approach to fast, online document clustering based on domain-specific similarity measures. Let $\mathbb{D}$ be the domain of documents d and $D \subset \mathbb{D}$ a given document collection, we define:

- a normalized document similarity measure S:

$$S \,:\, \mathbb{D} \times \mathbb{D} \rightarrow [0;1] \tag{1}$$

- a normalized similarity measure $\bar{S}$ between a set of documents and a single document:

$$\bar{S} \,:\, 2^{\mathbb{D}} \times \mathbb{D} \rightarrow [0;1] \tag{2}$$

$$\bar{S}(\bar{D}, \hat{d}) = \frac{\sum_{i=1}^{|\bar{D}|} S(d_i, \hat{d})}{|\bar{D}|} \tag{3}$$

Algorithm 1 ArteCM clustering algorithm

```
Require: Choose threshold parameter $\epsilon$
Require: Choose threshold parameter $\eta$
Require: Be $\hat{\mathbf{C}}$ a growing set of elements $C_i$ from $2^{\mathbb{D}}$
for all $d_j \in D$ do
    $m = \text{argmax}_i S(C_i, d_j)$
    if $\bar{S}(C_m, d_j) \geq \epsilon$ then
        if $\bar{S}(C_m, d_j) \leq \eta$ then
            $C_m = C_m \cup \{d_j\}$
        end if
    else
        $C^{new} = \{d_j\}$
        $\hat{\mathbf{C}} = \hat{\mathbf{C}} \cup C^{new}$
    end if
end for
```

The clustering algorithm we call ArteCM (see algorithm 1) requires the user to set two parameters:

1. a threshold parameter $\epsilon \in (0;1]$ that defines the minimum similarity $\bar{S}(C_i, d_j)$ a document d_j must have in order to be assigned to cluster C_i.

2. a threshold parameter $\eta \in [\epsilon;1]$ that defines the maximum similarity $\bar{S}(C_i, d_j)$ a document d_j must have to contribute to the definition of cluster C_i.

The two parameters play a fondamental role in the cluster growing process: the ϵ parameter directly controls the granularity of the document collection partitioning; while the η parameter controls the number of elements considered in similarity computations, having a strong impact on overall speed.

Two similarity measures are considered in our model:

1. a standard similarity measure S_D - the *Dice coefficent* [12] with binary *term weights*, appropriate for our context and defined as:

$$S_D(d_i, d_j) = \frac{2\mathcal{C}}{\mathcal{A} + \mathcal{B}} \tag{4}$$

where $\mathcal{C}$ is the number of common terms between d_i and d_j, $\mathcal{A}$ and $\mathcal{B}$ are the number of terms of d_i and d_j, respectively.

2. a novel similarity measure S_T aimed to better fit the nature of the short documents domain where a "weighted" similarity measure can be easily applied in which common terms contribute with different weights in function of their typology (numbers, words, special chars, ...).

$$S_T(d_i, d_j) = \sum_{r=1}^{|R|} \alpha_r \cdot \frac{2\mathcal{C}_r}{\mathcal{A}_r + \mathcal{B}_r} \tag{5}$$

such that $\sum_{r=1}^{|R|} \alpha_r = 1$ and where $F = \{f_1, \ldots, f_{|F|}\}$ is the set of term types and $\mathcal{C}_r$ is the number of common terms of type f_r between d_i and d_j, $\mathcal{A}_r$ and $\mathcal{B}_r$ are the number of terms of type f_r in d_i and d_j respectively.

3. Evaluation Metrics

The evaluation phase takes into account cluster quality and speed, since we want to investigate fast clustering algorithms that can be applied on the fly on a collection of documents.

In the Information Retrieval and Document Analysis field a widely accepted evaluation metric is the F-Measure (F1), as an armonic mean between Precision and Recall [3] indexes.

Given a collection of documents $D = \{d_1, \ldots, d_N\}$ and a list of labels $L = \{l_1, \ldots, l_M\}$ where $M \leq N$ we define the truth cluster set $\mathbf{C} = \{C_1, \ldots, C_M\}$ where $C_i = \{d_j : \text{the label of document } d_j \text{ is } l_i\}$.

If a single cluster C_i and an approximation of it $\hat{C}_j$ are considered, F_c is the F1 computed considering C_i as the set of relevant documents and $\hat{C}_j$ as the set of retrieved documents.

$$\text{F1}(\hat{\mathbf{C}}, \mathbf{C}) = \frac{\sum_{j=1}^{|\mathbf{C}|} |C_j| \cdot F1_c(\hat{C}_{\text{argmax}_i(F1(\hat{C}_i, C_j))}, C_j)}{\sum_j |C_j|} \tag{6}$$

Being $\hat{C}$ a cluster set computed by an algorithm and following [4] the $F1$ within two cluster sets can be computed in terms of $F1_c$: for each truth cluster the one with higher $F1$ is selected and then the weighted mean of $F1$ within all the cluster set is computed.

Table 1. Sample documents from PDA dataset

Document	Label
I-MATE Smartphone SP3 Windows	i-mate sp3
IMATE I-MATE SP3 TRI BAND BLUETOOTH ITAL	i-mate sp3
Palm Z22 32Mb Palm os 5.4	palm z22
Palm Z22 - Palm OS Garnet 5.4 200 MHz - ROM: 32 MB STN (160 x 160) - IrDA	palm z22

4. Experiments

Two experimental datasets [1] from the web commercial offer domain were used: PDAs and cell-phones, containing 3835 and 1218 documents respectively. Each dataset is composed of short descriptions of web product offerings, with supervised textual labels that permit to identify the *truth cluster set* used for evalution purposes: 90 and 332 truth clusters are present in the PDAs and cell-phones dataset, respectively. In table 1 sample documents and corresponding labels are reported.

Two configurations of ArteCM are considered in the experiments distinguished by the similarity measures adopted, S_D and S_T. In particular the S_T measure is defined with $F = \{f_1, f_2\}$ being f_1 the "word" term type and f_2 the "number" term type. Each configuration was varied setting parameters ϵ and η with different values; in particular the three best settings for each configuration are reported in tables 2 and 3. Both the α_1 and α_2 coefficients of the similarity measure S_T were set to 0.5 as this setting showed the most balanced behavior after a trial and error phase.

Three flat clustering techniques are compared with the proposed ArteCM model: a standard SinglePass implementation, an iterative K-Means (with randomly selected cluster seeds) and a SOM in its original version; all of them manage documents represented with standard VSM with boolean terms weighting scheme that in our trials reported the best results. Moreover, the Hierarchical Agglomerative Clustering Method has been evaluated by making clusters flat using a cutting threshold ϵ; the same framework of ArteCM has been setup, with a direct-text representation and similarity measures S_D and S_T. Both single link and complete link variants are considered. Results obtained[2] using PDAs and cell-phones respectively are showed in tables 2 and 3.

ArteCM clustering quality is in line or superior with other approaches, and an optimized balancing between speed and clustering quality can be found. In comparison with the standard SinglePass algorithm, the one with the best results together with the proposed approach, it can benefit from domain-tailored similarity measures. The number of clusters computed on average is about 200 for dataset PDA, and 350 for dataset cell-phones.

The K-Means iterative algorithm is able to provide quite good results, even though the need to know something about the number of needed clusters can be a limit in the web domain. Computing time though linear in the document collection size, can increase unexpectedly.

The HACM seems not to be very competitive with the employed direct-text representation, since it contrains to perform more costly similarity computations (in respect to, for example, dot product) and to choose only among single link and complete link merging schemes. The best performance was achieved with the single link, even though a different text representation permitting group averaging could lead to a more competitive scenario.

The SOM clearly shows its limits in terms of computing times and as a side effect of the F1 index with the increasing size of the document collection.

5. Conclusion

In this work short document clustering was addressed, proposing a novel strategy whose salient aspects are: the use of a flat representation-independent algorithm and a novel text-oriented similarity measure. As seen in the experiments focused on web commercial offerings, the overall strategy permits to reach high clustering quality in good balance with short computing times. Future works involve the investigation of the approach in the broader document clustering domain.

References

[1] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[2] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. In *WWW*

[1] Document collection taken from the price comparison service Shoppydoo (`http://www.shoppydoo.it`), and available on `http://www.dicom.uninsubria.it/arteLab/ricerca.html`

[2] The test configuration is an Ubuntu Linux 7.10 PC with Intel Core 2 CPU at 1,83 Ghz and 1 Gb of RAM

Table 2. F-Measures (F1) and time figures with the PDA dataset.

	F1 (*100)	Time (s)
ArteCM w/ S_T ($\epsilon = 0.5, \eta = 0.6$)	77.80	5.75
ArteCM w/ S_T ($\epsilon = 0.5, \eta = 0.7$)	78.35	10.18
ArteCM w/ S_T ($\epsilon = 0.5, \eta = 0.9$)	79.17	17.66
ArteCM w/ S_D ($\epsilon = 0.5, \eta = 0.5$)	59.74	2.84
ArteCM w/ S_D ($\epsilon = 0.5, \eta = 0.8$)	65.54	9.25
ArteCM w/ S_D ($\epsilon = 0.5, \eta = 1.0$)	64.99	10.55
SinglePass ($S_t = 0.5$)	66.77	3.06
SinglePass ($S_t = 0.6$)	64.44	4.92
SinglePass ($S_t = 0.7$)	55.53	7.03
HACM, SL, S_T ($\epsilon = 06$)	55.49	116.49
HACM, SL, S_D ($\epsilon = 0.7$)	43.40	111.94
K-Means ($K = 50$)	62.65	59.20
K-Means ($K = 90$)	63.46	57.68
SOM (8x8 map, $\sigma = 0.2$)	59.55	202.94
SOM (10x10 map, $\sigma = 0.2$)	62.50	317.19
SOM (15x15 map, $\sigma = 0.2$)	61.28	691.43

Table 3. F-Measures (F1) and time figures with the Cell-phones dataset.

	F1 (*100)	Time (s)
ArteCM w/ S_T ($\epsilon = 0.5, \eta = 0.5$)	77.40	12.00
ArteCM w/ S_T ($\epsilon = 0.5, \eta = 0.7$)	80.31	54.21
ArteCM w/ S_T ($\epsilon = 0.5, \eta = 0.9$)	82.22	121.76
ArteCM w/ S_D ($\epsilon = 0.5, \eta = 0.6$)	60.86	41.31
ArteCM w/ S_D ($\epsilon = 0.5, \eta = 0.7$)	68.65	78.99
ArteCM w/ S_D ($\epsilon = 0.5, \eta = 0.8$)	72.01	97.89
SinglePass ($S_t = 0.5$)	57.91	19.97
SinglePass ($S_t = 0.6$)	65.86	36.80
SinglePass ($S_t = 0.7$)	61.08	75.52
HACM, SL, S_T ($\epsilon = 0.6$)	62.75	3452.76
HACM, SL, S_D ($\epsilon = 0.6$)	49.40	3462.60
K-Means ($K = 100$)	41.48	506.55
K-Means ($K = 332$)	55.51	1274.81
SOM (8x8 map, $\sigma = 0.2$)	32.57	932.17
SOM (10x10 map, $\sigma = 0.2$)	38.65	1507.98
SOM (15x15 map, $\sigma = 0.2$)	48.16	3415.88

'05: Special interest tracks and posters of the 14th international conference on World Wide Web, pages 801–810, New York, NY, USA, 2005. ACM.

[3] W. B. Frakes and R. A. Baeza-Yates, editors. *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall, 1992.

[4] K. Hammouda and M. Kamel. Efficient phrase-based document indexing for web document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 16(10):1279 – 1296, 2004.

[5] J. Hartigan and M. Wang. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[6] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, volume 1, pages 413–418. IEEE Service Center, Piscataway, NJ, 1998.

[7] K. Lagus, S. Kaski, and T. Kohonen. Mining massive document collections by the websom method. *Inf. Sci.*, 163(1-3):135–156, 2004.

[8] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings 5th Berkeley Symposium*, pages 281–297, 1967.

[9] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the second international conference on Human Language Technology Research*, pages 280–285, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[10] D. A. Metzler Jr. *Beyond Bags of Words: Effectively Modeling Dependence and Features in Information Retrieval*. PhD thesis, University of Massachusetts Amherst, 2007.

[11] G. Salton. *The SMART Retrieval System*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[12] G. Salton, editor. *Automatic text processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1988.

[13] G. Salton and M. J. Mcgill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[14] R. Sharma and S. Raman. Phrase-based text representation for managing the web documents. In *ITCC '03: Proceedings of the International Conference on Information Technology: Computers and Communications*, page 165, Washington, DC, USA, 2003. IEEE Computer Society.

[15] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques, 2000.

[16] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based serach engine for complex queries. 2004.

[17] B. Tang, M. Shepherd, M. I. Heywood, and X. Luo. *Comparing Dimension Reduction Techniques for Document Clustering*, pages 292–296. Springer Berlin / Heidelberg, 2005.

[18] T.Kohonen. *Self-Organizing Maps*. Springer-Verlag New York, Inc., 1995.