

Clustering-Based Relevance Feedback for Web Pages

Seung Yeol Yoo and Achim Hoffmann

School of Computer Science and Engineering
University of New South Wales Sydney 2052, NSW, Australia
{syy, achim}@cse.unsw.edu.au

Abstract. Most of traditional relevance feedback systems simply choose top ranked Web pages for a query as the source of providing the weights of candidate query expansion terms for the query. However, the whole contents of such top-ranked Web pages are usually mixed with sub-topically distinguishable contents that are too heterogeneous to be directly used to extract good quality candidate query expansion terms.

In this paper, our basic idea is that the Web pages properly clustered into a sub-topic cluster can be used as a better source rather than whole given Web pages, to provide more topically coherent relevance feedback for that specific sub-topic. Thus we proposed **Clustering-Based Relevance Feedback for Web Pages**, which utilizes three methods to cluster given or retrieved Web pages into several subtopic-clusters. These three methods cooperates to construct good quality clusters by respectively supporting Web pages Segmentation, Terms (or Features) Selection, k Seed Centroids Selection. Here the automatically selected terms indicate the relevance feedback to construct all sub-topic clusters and assign the given Web pages into proper clusters. Each subset of the selected terms, which occurs in the Web pages assigned into a sub-topic cluster, indicates the relevance feedback to expand a query over that sub-topic cluster. Our experimental results showed that the clustering performances based on two traditional term-weighting methods (i.e., an unsupervised method and a supervised method) were significantly improved with our methods.

1 Introduction

While the Web represents a vast information resource, the entire contents in each Web page might not always be valuable to support a user's specific information needs. On the same Web page, the author of the Web page usually puts a lot of contents, which belong to different subtopics [1]. Thus, the content should be disassembled into smaller elements (e.g., segments) than Web pages. Such finer-grained elements should be well organized to be properly accessible according to a user's specific information needs.

To achieve such purposes, we need to consider the commonality and the difference between two views; the author's view (i.e., intention) described in a Web page and a user's view (i.e., attention) of the Web page. For example, in (a) of Figure 1, a user similarly recognized the topical streams intended by the author of a Web page, but the topical importance value of each segment was different to the user from that intended by the author. We call the author's view 'objective basic view' and a user's view 'subjective query view'. The 'subjective

to generate with given Web pages, and 2) the user-selections of expansion terms are simulated by an unsupervised or a supervised term selection method.

The rest of this paper is organized as follows: Related works are presented in Section 2. Section 3 describes our approach utilizing three methods which select more coherent expansion terms from the given Web pages and then utilize the selected expansion terms for constructing Web pages clusters. Section 4 describes the set up of our experiments. Section 5 presents our experimental results showing that the average quality of clusters constructed by our approach is better than that of clusters constructed by k-means clustering based on two traditional terms selection methods. Section 6 concludes this paper with future work.

2 Related Works

As noise terms and multi-topics are two major negative factors for expansion terms selection in a relevance feedback technique, it is necessary to *segment a web page into semantically related units* so that irrelevant information and misleading terms can be filtered out and multiple topics can be distinguished. For example, [1–6] using segmentation method showed that it is necessary to filter out noisy segments and detect important segments from Web pages for the improvement of web information retrieval. However, their works did not touch the second negative factor to provide focusing on specific subtopics within Web pages by considering topically distinguishable boundaries within Web pages.

The Vision-based Page Segmentation (VIPS) method [7] is suggested to supplement the Dom-based segmentation method³ e.g., [4] with visual cues. For example, [6] works in this way: First they calculated the relevance values among content-elements according to their visual adjacency and/or appearance similarity. And then, they applied such relevance values to a few heuristic rules (i.e., a set of rules identifying the neighborhood relationship among content-element) to group relevant content-elements in a Web page into several segments.

[6] expected some visual cues can aggregate more semantic information which are difficult to directly extract from html-tags within the html source of a Web page. However, visual cues have some problems such as 1) how to choose the proper size of the unit content-element and decide generic adjacency-conditions between two content-elements, which affect the grouping of content-elements in the same segment. For example, the adjacency between two content-elements can be affected by the size of intermediate content-element(s), and 2) the decision about the total number of visually distinguishable segments gives significant effects to the sizes of content-elements and the interpretation of those content-elements' semantic relationships. However such decisions cannot be predefined without considering the content of each Web page, and no satisfactory solution has been found as yet.

Another criticism of the VIPS method for discovering segments and their structure are as following: it does not distinguish two segments which describing different subtopics in similar layouts (e.g., similar size, length and position). According to our observations on Web pages, segments enumerating different

³ It parses a Web page into a tag tree, based on Html-tags' syntactic structure, to reveal the presentation structure of the Web page.

subtopics are frequently presented with similar layouts in a Web page (e.g., in a portal site). Thus visual cues are too coarse a method to detect subtopic-specific segments. For example, topically different contents published with similar visual cues are considered to have the same importance value in [5], without considering the topical differences of the contents.

In this respect, the most similar system to ours was introduced in [1], because they tried to understand the hierarchical topic-structure which is implicitly described within the content of a Web page and the various information needs of a user by inferring the user’s browsing behaviors. However, this approach did not answer the following two problems: Firstly, the distinctions among the user’s browsing behaviors, performed on the same Web pages for different topic interests at different times, are not considered. Secondly, the tracking of user’s behaviors is too a time-consuming task to detect user’s interests at query run-time.

3 Relevance Feedback based on Web Pages Clustering

3.1 Layout-Based Segmentation

The goal of layout-based segmentation of a Web page is to discover more precise narration structure intended by the author of the Web page.

With those reasons mention in previous section, we selected a way extracting Generic Nested-Structure Patterns (GNSPs) rather than utilizing frail visual cues from a html-source to interpret the structural semantics in the html-source. Our GNSPs extraction algorithm is a way supplementing the limitations [6] of the Dom-based segmentation method with nest-based hierarchy relationships between html-tags. We consider that the narration structure intended by an author can be more precisely extracted with the help of GNSPs rather than visual cues, because GNSPs can naturally: 1) bound each content-element according to the html-tags formatting it (Such boundary is not affected by the change of a content-element’s size), 2) construct segments-topic-paths identifying the adjacency between content-elements, and 3) determine the number of distinguishable segments and segments-topic-paths of a Web page.

For example, Figure 2 illustrates the topic narration structure among hierarchically connected segments of a Web page, which are extracted by our GNSPs extraction algorithm. Here, the segment **C** contains five sub-segments **C1**, **C2**, **C3**, **C4** and **C5**. Within the html-source, the sub-segments are grouped with “<table>” tag, each sub-segment is separated with “<div>” tags and each segment’s title (e.g., **C1** has title “November 18, 2004”) is modified with “” tag. According to our observations, “” is usually used to modify the characteristics of a particular text (e.g., title) and “<div>” is intended for group-level formatting.

GNSPs extraction algorithm was performed as a query pre-processing task not to obstruct the run-time performance of a information retrieval system as follows: 1) remove noisy contents such as script-codes, 2) analyze the generic hierarchy structure relationships among html-tags. For example, the listing information (e.g., , , , <select>, <option>, etc.), the block information (e.g., <table>, <tr>, <td>, etc.), the boundary information (e.g., <hr>, <h#>, <thead>, <tbody>, <tfoot>, <div>, <p>,
, etc.), and the priority information (i.e., some tags should be considered with higher priority than

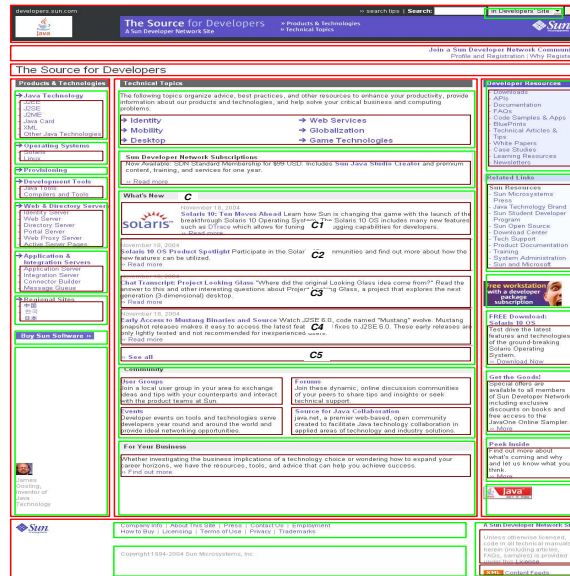


Fig. 2. The discovered narration structure of the Web page at <http://developers.sun.com/>. The hierarchy relationships between segments are illustrated by nesting and nested color-rectangles.

the other tags. E.g., a `<div>` tag occurring between `<table>` and `</table>` tags should separate only the content between `<table>` tag and `</table>` tag rather than the whole contents of the Web page).

We notice that more details of GNSPs extraction algorithm is omitted with the limited space of this paper, and the evaluation of this algorithm is replaced with the evaluations of clusters constructed by applying this algorithm to the clustering tasks.

3.2 Term Selection Methods

The performance of clustering task depends on the quality of terms extracted with a terms selection method. In this subsection, we would introduce two traditional methods (χ statistic method (CHI) and Term Contribution method (TC)), and propose a new method (Term Context Contribution based on Segments-Topic-Paths (TCC-STP)).

χ^2 Statistic (CHI) The χ^2 statistic is a supervised method which measures the statistical significance of association between a term and a category [8] by using known class label information. The statistical significance can be defined to be $\chi^2(t, c) = \frac{N \times (p(t, c) \times p(\bar{t}, \bar{c}) - p(t, \bar{c}) \times p(\bar{t}, c))}{p(t) \times p(\bar{t}) \times p(c) \times p(\bar{c})}$ and $\chi^2(t) = \text{avg}_{i=1}^m \{\chi^2(t, c_i)\}$.

Term Contribution (TC) Term Contribution method (TC), proposed as an unsupervised method in [9], considers the different importance of a term in

different documents. They define the contribution of a term in a dataset as its overall contribution to the documents’ similarities. The definition for TC is $TC(t) = \sum_{i,j \cap i \neq j} f(t, w_i) \times f(t, w_j)$, where $f(t, w)$ represents the $tf * idf$ [10]. However TC does not consider different contexts of each term in different documents. For example, if a term “java” is used with terms “island” and “travel” in a document, and is used with terms “software” and “applet” in another different document, then the contribution of “java” should be separately considered for two different contexts (i.e., two different word-senses “island” and “programming language” of “java”).

Term Context Contribution based on Segments-Topic-Paths (TCC-STP) To consider different contexts of a term in different Web pages, we introduce a new unsupervised feature selection method called “Term Context Contribution based on Segments-Topic-Paths” (TCC-STP). It gives higher weights to the terms sharing more common context-words within the same segments-topic-paths. The equation for TCC-STP is:

$$TCC - STP(t) = \sum_{i,j \cap i \neq j} f(t, w_i) \times f(t, w_j) \times f(CT(i, j), STP(t, w_i)) \times f(CT(i, j), STP(t, w_j))$$

, where $STP(t, w)$ represents segments-topic-paths, in which t is occurred, of Web page w . $CT(i, j)$ denotes context-words co-occurring in $STP(t, w_i)$ and $STP(t, w_j)$. $f(CT(i, j), STP(t, w_i))$ represents the frequency of $CT(i, j)$ in $STP(t, w_i)$. We simply set $f(CT(i, j), STP(t, w_i))=1$ and $f(CT(i, j), STP(t, w_j))=1$, when $|CT(i, j)|=0$

3.3 Applied K-means Clustering (AKC)

K-means Clustering (**KC**) method is an iterative approach starting with randomly selected k centroids and assigning each Web page to its nearest centroid(s)⁴. Then, iteratively, based on the assigned Web pages to each cluster, the respective centroids are updated and Web pages are re-assigned to the nearest centroids.

To select more coherent expansion terms, we modified the original k-means clustering method with following two algorithms.

Initial k Centroids Selection In k-means clustering, the number of iterations and the quality of the final clusters depend on choosing proper number k of centroids and selecting good initial centroids (i.e., seed centroids). For experiment purposes, we assumed that k is the same with the number of original categories within test-dataset.

We selected k seed centroids in a way which maximizes the total distance of the selected seed centroids. For example, in Figure 3 (there, let us assume

⁴ With a “soft clustering” method, a Web page can be assigned to multiple centroids. However, in this paper, we followed a “hard clustering” method assigning each Web page to only one centroid.

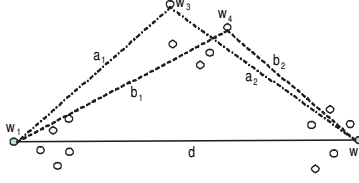


Fig. 3. An example of centroid selections. d , a_1 , a_2 , b_1 and b_2 respectively represents the similarity distance between two Web pages.

$k=3$), the longest distance between Web pages is d , thus we selected Web pages w_1 and w_2 as two seed centroids. And then, a Web page which is the most far from previously selected seed centroids is selected as the third seed centroid. For instance, in Figure 3, $\sum_i a_i > \sum_j b_j$ thus w_3 is selected as the third seed centroid.

We denote this algorithm as a maximization function of total inter-centroids distance D :

$$Max D = \sum_{i=1}^{k-1} \sum_{j=i+1}^k \overline{w_{c_i} w_{c_j}}$$

where w_{c_i} and w_{c_j} respectively represents a Web page selected as the seed centroid for cluster c_i and c_j . The distance $\overline{w_{c_i} w_{c_j}}$ between two seed centroids can be

calculated by one of the following two equations: 1) $\overline{w_{c_i} w_{c_j}} = -\sum_{t \in T} \left\{ \frac{(f(t, w_{c_i}) + f(t, w_{c_j}))}{2} \times \frac{|T_m|}{|T|} \right\}$ if $f(t, w_{c_i}) \neq 0$ and $f(t, w_{c_j}) \neq 0$, and 2) $\overline{w_{c_i} w_{c_j}} = \sum_{t \in T} \left\{ \max(f(t, w_{c_i}), f(t, w_{c_j})) \times \frac{|T_n|}{|T|} \right\}$ if $(f(t, w_{c_i}) = 0 \text{ and } f(t, w_{c_j}) \neq 0)$ or $(f(t, w_{c_i}) \neq 0 \text{ and } f(t, w_{c_j}) = 0)$.

Where T denotes all terms occurred in w_{c_i} or w_{c_j} . $f(t, w_{c_i})$ is the occurrence frequency of term t in w_{c_i} . $T_m \subset T$ and $T_n \subset T$ is the terms respectively satisfying the if condition of those equations. $\frac{|T_m|}{|T|}$ and $\frac{|T_n|}{|T|}$ is used to normalize the frequency weight of each term t over the total number $|T|$ of terms occurred within two Web pages, and give more weight to common terms and distinguishable terms, respectively. The former $\overline{w_{c_i} w_{c_j}}$ equation reduces the distance between two seeds when they share a term t , but the later $\overline{w_{c_i} w_{c_j}}$ equation increases the distance between two seed centroids when a term t uniquely occurs in only one centroid w_{c_i} or w_{c_j} .

Our centroids selection method has following advantages: 1) the number of clustering iterations until convergency can be reduced and the quality of the final clusters can be improved and 2) there is no need to average several times the k -means clustering results based on different sets of centroids because our initial centroids selection algorithm returned always more optimized results through our experimental evaluation (Compare the performance measures of *-AKC and *-KC, in Figure 4).

Similarity with a Seed Centroid When k seed centroids are decided, the clustering problem can be considered as a classification problem which assigns each Web page to the most similar seed centroid among the k seed centroids.

The given Web pages $W \ni w_i$, the terms $T \ni t$ occurred in W and the occurrence frequencies $f(w_i, t)$ form a vector space. With that vector space, traditional k-means clustering method tried to minimize total intra-cluster distance. Each intra-cluster distance is measured by the sum of terms' occurrence differences between a centroid and assigned Web pages to that centroid. Thus, total intra-cluster distance can be denoted as a following equation: $Min Distance = \sum_{j=1}^k \sum_{w_i \in W_{c_j}} |w_i - c_j|$, where there are k clusters W_{c_j} , $j = 1, \dots, k$. c_j denotes the centroid (i.e., mean point) of the Web pages W_{c_j} .

However, for Web pages, the traditionally used euclidian-distance measure is not efficient to cluster relevant Web pages. One reason is that common terms are usually much less than different terms between a Web page w_i and a seed centroid c_j , even they talk about common subtopic(s). Thus, there is always a chance of that a Web page sharing a few common terms but also having huge different terms would have large distance value than another Web page not sharing any common terms but having very few different terms with the centroid. As the result, two Web pages assigned into the same centroid may be considered to be relevant to each other, because they have less different terms rather than more common terms with the centroid. Another reason is that the euclidian-distance makes even such rarely common terms increase the distance according to the difference of their occurrence frequencies.

To answer such issues in Web pages clustering, we reformalized the distance minimization problem between centroids and given Web pages as a similarity S maximization problem.

$$Max S = \sum_{c_j=1}^k \sum_{w_i \in W_{c_j}} \{S'(w_i, c_j) - D'(w_i, c_j)\}$$

We define the similarity between a centroid c_j and a Web page w_i as the subtraction of Dissimilarity from Similarity. Similarity and Dissimilarity can be denoted as following two functions: $S'(w_i, c_j) = \sum_{t \in T} \{(f(t, w_i) + f(t, c_j)) \times \frac{|NZ(c_j, T)|}{|T|}\}$, if $f(t, c_j) \neq 0$ and $t \notin NZ(\bar{c}_j, T)$. $D'(w_i, c_j) = \sum_{t \in T} \{f(t, w_i) \times \frac{|Z(c_j, T)|}{\max(1, |NZ(\bar{c}_j, T)|)}\}$, if $f(t, c_j) = 0$ and $t \in NZ(\bar{c}_j, T)$. Where $f(t, w_i)$ and $f(t, c_j)$ respectively denotes the occurrence frequency of term t in the Web page w_i and the centroid c_j . T denotes a set of terms occurred in a Web page w_i . $NZ(c_j, T) \subset T$ denotes the terms appearing in both w_i and c_j . $Z(c_j, T) \subset T$ denotes the terms appearing in w_i but not in c_j . $NZ(\bar{c}_j, T) \subset Z(c_j, T)$ denotes the terms appearing in w_i and at least one of the other centroids excluding c_j .

4 Evaluation Configuration

Through the experiments, our goal is to show that TCC-STP and AKC methods mentioned in Section 3 can extract good quality query expansion terms which can indicate relevant Web pages. For this purpose, the most intuitive evaluation method might be to compare the expansion terms extracted by a system with those extracted by human subjects. However, it has some issues like that the consistency problem between different human subjects, the limited human analysis capacity and etc. Thus, we followed a reverse but automatic way comparing

the qualities of obtained clusters by different query expansion term extraction methods. Better clustering performance can be achieved by using better query expansion terms and their weights as the clustering features and feature values.

DataSet	Topic Num	Docs Num	Dis. Words	Avg. Dis. Words	Avg. DF per Dis. Word
A	5	96	6,010	63	3
B	21	333	13,833	42	6
C	12	369	13,031	35	6

Table 1. The dataset properties. *Dis.words* means stemmed words (i.e., lemmas). DF means the number of documents in which a *Dis.word* occurred.

4.1 Standard Clusters for Web Pages

To test the clustering performance, we collected three datasets from Yahoo! Directory (Downloading prohibited, ASP, link-broken or multi-framed Web pages are removed from each category). The information about the datasets is shown in Table 1. Dataset A, B and C is respectively relevant to “Robotic”, “Computer Science” and “Computer and Internet” category in Yahoo! Directory, and respectively includes 5, 21 and 12 subcategories. We considered each subcategory as one standard cluster, in the aspect of main topics occurred within the three datasets.

4.2 Evaluation Measures

Entropy and F-measure were used to evaluate the clustering performance. Those measures of a cluster respectively indicates the uniformity and the weighted harmonic mean of precision and recall of assigned Web pages to the cluster.

The Entropy of all obtained clusters is defined by the weighted sum of the entropy of each obtained cluster, as following: $Entropy = \sum_{k=1}^{C'} \frac{|W_k|}{N} \sum_{j=1}^C p_{jk} \times \log(p_{jk})$, where $p_{jk} = \frac{1}{|W_k|} |\{w_i | label(w_i) = c_j\}|$ is the entropy of a obtained cluster. N is the total number of Web pages in the tested original clusters. C' , C denote the number of obtained clusters and the number of original clusters respectively. W_k is the set of Web pages in a obtained cluster. $label(w_i)$ returns the cluster-label of the Web page w_i . The cluster-label of a generated cluster is decided by the original cluster-label of the most shared Web pages in the generated cluster.

The Precision of a obtained cluster $W(\exists w_i, i = 1, \dots, |W|)$ is defined as following: $Precision(W) = \frac{1}{|W|} \max(|\{w_i | label(w_i) = c_j\}|)$. The Recall of W is defined as following: $Recall(W) = \frac{1}{|O|} \max(|\{w_i | label(w_i) = c_j\}|)$, where $label(O) = c_j$. $|O|$ denotes the number of Web pages which have cluster-label c_j , where $O \subset W$. The traditional F-measure (i.e., F1 measure; recall and precision are evenly weighted) of a obtained cluster is defined as following: $F(W) = \frac{2 \times Precision(W) \times Recall(W)}{Precision(W) + Recall(W)}$. The sum of F-measures of all obtained clusters C' were averaged by the number of all obtained clusters.

5 Experimental Results and Discussion

We performed clustering tasks over two test-datasets T1 and T2 with 4 different configurations and 8 different terms selection rates. The test-dataset T1 has 10 sub-datasets, and each sub-dataset consists of randomly selected one category from each original dataset A, B and C, and is composed of average 50 Web pages. The test-dataset T2 has 10 sub-datasets and each sub-dataset consists of from 3 to 7 categories which have ancestor and descendant relationship each other within a original dataset A, B or C, and is composed of 84 Web pages.

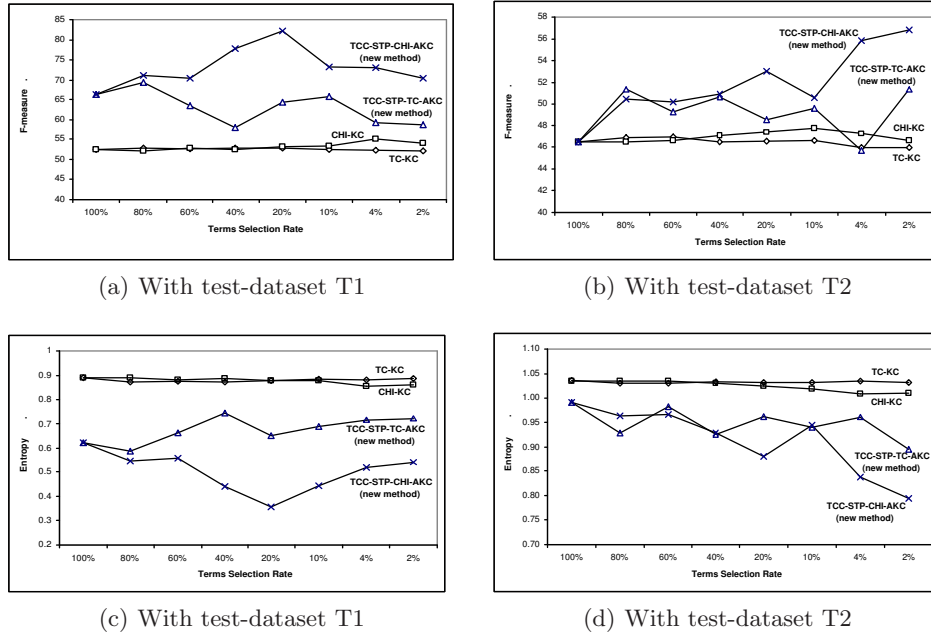


Fig. 4. Average F-measures and Average Entropies of clusters, which are achieved with different top-ranked query expansion terms

The tested four different configurations are named as following: **TC-KC**(Term Contribution(TC) + K-means Clustering(KC)), **CHI-KC**(χ^2 statistic(CHI) + K-means Clustering(KC)), **TCC-STP-TC-AKC**(Term Context Contribution based on Segments-Topic-Paths(TCC-STP) + Term Contribution(TC) + Applied k-Means Clustering(AKC)), and **TCC-STP-CHI-AKC**(Term Context Contribution based on Segments-Topic-Paths(TCC-STP) + χ^2 statistic(CHI) + Applied K-means Clustering(AKC)). Our methods TCC-STP and AKC (explained in subsection 3.2 and 3.3) are applied to TC-KC and CHI-KC configurations.

To consider the variances of K-means clustering (KC) influenced by the initially selected centroids, we randomly produced 10 sets of initial centroids for each tested sub-dataset within T1 and T2. Then averaged 10 times the clustering performances to calculate the final clustering performance of a tested

sub-dataset. Comparatively, the initial k centroids of Applied K-means Clustering (AKC) were decided as k Web pages which maximize the distances between them, without any needs to consider the variances of clustering results (See subsection 3.3 for more details).

(a) and (b), and (c) and (d) in Figure 4 respectively show the F-measure values and Entropy values of obtained clusters over test-dataset T1 and T2. The average clustering performance based on two traditional term-weighting methods the unsupervised method TC and the supervised method CHI was respectively improved by 21%(from 52% to 63%) and 41%(from 53% to 75%) for the F-measure and by 24%(from 0.88 to 0.67) and 48%(from 0.88 to 0.5) for the Entropy, over the test-dataset T1. Over the test-dataset T2, the average F-measure value improved by 6%(from 46% to 49%) and 8%(from 47% to 51%) for the F-measure and by 8%(from 1.03 to 0.95) and 12%(from 1.03 to 0.91) for the entropy. These improvements indicate that many noisy terms (e.g., high weighted for whole given Web pages but is not for the Web pages assigned to a specific cluster) were removed by our TCC-STP and AKC methods.

The overall performance values of four different configurations (TC-KC, CHI-KC, TCC-STP-TC-AKC and TCC-STP-CHI-AKC) were not so high. The reason for that might be the average ‘DF per Dis. word’ is very low even among the Web pages belong to the same original category. For example, the average ‘DF per Dis. word’ was respectively 3, 6, 6 for dataset A, B and C. It makes difficult for the system to cluster relevant Web pages, in the case of test-dataset T1. In the case of test-dataset T2, the average ‘DF per Dis. word’ became higher than that of test-dataset T1. It is because the original categories in each sub-dataset of T2 are topically much relevant each other (i.e., they have the ancestor and descendant category relationship each other). In this case, our ‘similarity calculation method between a seed centroid and a web page’ met a negative situation, since the similarity value might be too similar to distinguish a Web page originally assigned to a ancestor category from another Web page originally assigned to the child category of the ancestor category. In some cases, the similarity value between two Web pages in a original category was smaller than that of two Web pages which belong to different but topically adjacent categories. We consider that such similar Web pages might can be combined into the same category, if the category structure of Yahoo! is refined at the point of view focusing on the common topic of the similar Web pages.

With our new methods TCC-STP and AKC methods, the best performance is achieved by the TCC-STP-CHI-AKC configuration, and it might be due to the benefits of using known class label information to extract good quality query expansion terms. Even though TCC-STP-CHI-AKC and CHI-KC utilized the same class-label information for clustering tasks, TCC-STP-CHI-AKC showed better performance with our TCC-STP and AKC methods. We also would like to concentrate on the performance of the TCC-STP-TC-AKC configuration which get the second best performance. TCC-STP-TC-AKC is a new unsupervised method for term selection. In practice, the class labels of given Web pages cannot be known before any query sessions, thus TCC-STP-TC-AKC might be the best solution to extract good quality query expansion terms among those four configurations. The F-measure values and Entropy values of TCC-STP-TC-AKC outperformed those of CHI-KC which is based on the supervised method CHI.

6 Conclusion

In this paper, the experimental results showed that our methods can significantly improve the performance of clustering tasks based on the supervised term-weighting method CHI and the unsupervised method TC. When our methods TTC-STP (Term Context Contribution based on Segment-Topic-Paths; a new unsupervised term-weighting method) and AKC (a newly applied K-means clustering approach; Initial k Centroids Selection method and Similarity-based clustering method) cooperated to cluster relevant Web pages, it outperformed the performance of the traditional clustering method CHI-KC (based on a supervised term-weighting method).

Further developments of our approach will investigate the possibility of interactive query-expansion. More coherent and reduced query expansion terms can be extracted from the clustered Web pages into a cluster in which a user is interested. Iteratively, the user can select some of the query expansion terms to refine their query expression and then start a new clustering process over the clustered Web pages under the context of the refined query expression.

References

1. Yin Liu, W.L., Jiang, C.: User interest detection on web pages for building personalized information agent. In: WAIM 2004: the 5th international Conference on Web-Age Information Management. (2004) 280–290
2. Deng Cai, Shipeng Yu, J.R.W., Ma, W.Y.: Block-based web search. In: SIGIR-2004: the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2004) 456–463
3. Deng Cai, Shipeng Yu, J.R.W., Ma, W.Y.: Extracting content structure for web pages based on visual representation. In: Web Technologies and Applications, 5th Asian-Pacific Web Conference, APWeb 2003. (2003) 406–417
4. Suhit Gupta, Gail Kaiser, D.N., Grimm, P.: Dom-based content extraction of html documents. In: WWW 2003: the 12th International World Wide Web Conference. (2003) 207–214
5. Ruihua Song, Haifeng Liu, J.R.W., Ma, W.Y.: Learning block importance models for web pages. In: Proceedings of the 13th International World Wide Web Conference, WWW 2004. (2004) 203–211
6. Shipeng Yu, Deng Cai, J.R.W., Ma, W.Y.: Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In: Proceedings of the 12th International World Wide Web Conference, WWW2003. (2003) 11–18
7. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Vips: a vision-based page segmentation algorithm. Technical report, Microsoft (2003)
8. Luigi Galavotti, F.S., Simi, M.: Experiments on the feature selection and negative evidence in automated text categorization. In: Research and Advanced Technology for Digital Libraries, 4th European Conference (ECDL 2000). (2000) 59–68
9. Tao Liu, Shengping Liu, Z.C., Ma, W.Y.: An evaluation on feature selection for text clustering. In: ICML 2003: the Twentieth International Conference on Machine Learning. (2003) 488–495
10. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-wesley, Boston, MA, USA (1989)