

# Query-Topic Focused Web Pages Summarization

Seung Yeol Yoo and Achim Hoffmann

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

**Abstract.** We present a novel Web Pages Summarizer **ContextSummarizer** that subgroups the given Web pages into ‘sense-clusters’ respecting a user’s topic interests, and constructs a dynamic extractive summary for each sense-cluster. A user’s topic interest is described by the user who selects and refines some of word senses disambiguated within the content contexts of the given Web pages. The semantic similarity measures between the contents of Web pages/segments/sentences and the user-selected/-refined word senses were used to choose the most topically relevant sentences as the extractive summaries referring to a user’s topic interest. As the results, it addressed the dynamic semantic-alignment issues between the content of a Web page and the user’s topic interest about that Web page, and between the user’s topic interest and an extractive summary. Some case studies and experimental results showed that query-topic focused extractive summaries returns more topically consistent sentences for an extractive summary.

## 1 Introduction

The continuous and rapid growth of Web information requires efficient assistances to Web users for information filtering and interpretation. Such efficient assistances are expected to be helpful in elucidating the occurred topics within the Web pages retrieved by a Web information retrieval system (e.g., a keyword-based search engine), and organizing the retrieved Web pages into topically associated collections (e.g., a Web directory system such as Yahoo!). However, the topic-based elucidation and organization of multiple Web pages is difficult to be achieved without automated summarization techniques due to the labor-intensive and diverse nature of a user’s understanding over the contents of multiple Web pages. In this context, fully automatic summarization, having similar accuracy like the summary by a human who can extract the gist of Web information according to their information needs, is not yet possible in real world. For instance, even though general purpose summaries of multiple Web pages can be extracted by a fully automatic method (e.g., [1]), such summaries might be potentially noised by irrelevant information against a user’s specific topic interests.

In this paper, we propose a novel Human Aided Machine Summarization (HAMS) system, called **ContextSummarizer**. ContextSummarizer goes beyond previous HAMS systems such as [2] where a user should highlight passages

within a Web page. The user-highlighted passages were utilized as the source for the generation of topic-focused summary. With such system, a user has to observe not only a local view over each target Web page but a global view over whole target Web pages to properly choose the most topically interesting content segments, and thus it might necessary cause the user’s perception overburdens. For the reduction of a user’s perception overburdens related to the Web pages summarization tasks, our system iteratively suggests a set of classifiers (i.e., context-words and their senses) discovered from the target Web pages, and then asks the user to select only topically relevant classifiers to their topic interests (From now on, we interchangeably use a user’s topic interest(s) or topic attention(s) over a collection of Web pages with ‘*query-topic(s)*’). In other words, our principle is to reformulate the perception problem over the target Web pages as the recognition problem over a set of classifiers suggested by our system. According to cognitive psychology findings, perception is more demanding than recognition.

In this paper, we pointed one principle problem of generic summarization systems (e.g., Copernic Summarizer [3] and Pertinence Summarizer [4]): Without any topic focus for a summarization, we cannot even evaluate which extractive summary is better or not. To demonstrated the strengthes of ContextSummarizer, we also compared the query-topic focused ‘extractive summaries’ generated by our ContextSummarizer with those generated by Columbia Newsblaster [5]. The Columbia Newsblaster subgroups the given Web pages into a few predefined topic categories, and then respectively generates an extractive summary on the Web pages clustered into each topic category. As an extrinsic evaluation result, we showed the terms extracted from our extractive summaries improves the performance of Web pages clustering.

This paper is organized as follows: In Section 2, we explain about some related works. Section 3 briefly describes the system architecture composed with ContextExplicator, ContextSegmentor and ContextSummarizer. In Section 4, we introduce our summarization algorithm. In Section 5, we presents our initial evaluation results based on some case studies and an extrinsic evaluation. Finally, the conclusion is presented in Section 6.

## 2 Related Works

Recently much work has been done on Web Page(s) summarization (e.g., [1, 2, 6–12]). In those works, Web pages are generally considered as having quite different characteristics in both structure and content, compared to pure-text documents. For example, as the difficulty of extracting well-defined and coherent discourse structures, [6, 7] followed an abstractive summarization approach rather than an extractive summarization approach. As the difficulty of explicitly representing a query-topic and aligning the semantics between given Web pages and the query-topic, most of extractive summarization systems focused on constructing generic summaries rather than topic-focused summaries.

To answer those two kinds of difficulty, we consider that a sophisticated Web pages summarization system should be supported by following two functions:

First, it discovers the implicitly embedded discourse structures in Web pages. Relevant contents within the Web pages would be bounded by the discovered discourse structure information, and then the contextual features occurred within relevant contents can provide more coherent information for Web pages summarization. Second it starts a summarization process with the construction of a user’s query-topic rather than assuming that the main topic of the Web pages is always the same with a user’s current query-topic. It is because Web users might want to get the summary of Web pages that are relevant to their topic interests rather than the most commonly conveyed topics across the Web pages. Thus the topically relevant Web pages among the originally given Web pages and the topically relevant contents (e.g. content segments) within the topically relevant Web pages should be discovered according to a user’s specific topic interests. More topically coherent source, matched with a specific query-topic, would make it possible for a Web pages summarizer to construct more probabilistically and/or statistically qualified extractive summary for that query-topic.

Copernic Summarizer [3] uses statistical and linguistic algorithms for single-document summarization. This system pinpoints “key concepts” to sub-select conceptually (i.e., topically) relevant sentences as a generic summary. However, the “key concept” is limited to work as a term (i.e., a term does not distinguish possibly different senses appearing on different sentences and documents) rather than a concept (i.e., a sense of the term, which can be distinguished from the other senses of the term) for summarization purpose. Thus such “key concept” cannot be applied for multiple Web pages summarization that necessary requests for a summarizer to distinguish different senses of a term occurring within Web pages. In this aspect, Newsblaster [5] has shown a good starting point by grouping articles together into predefined topic categories and summarizing the clustered Web pages into each topic category.

## 2.1 Our Approach for Web Pages Summarization

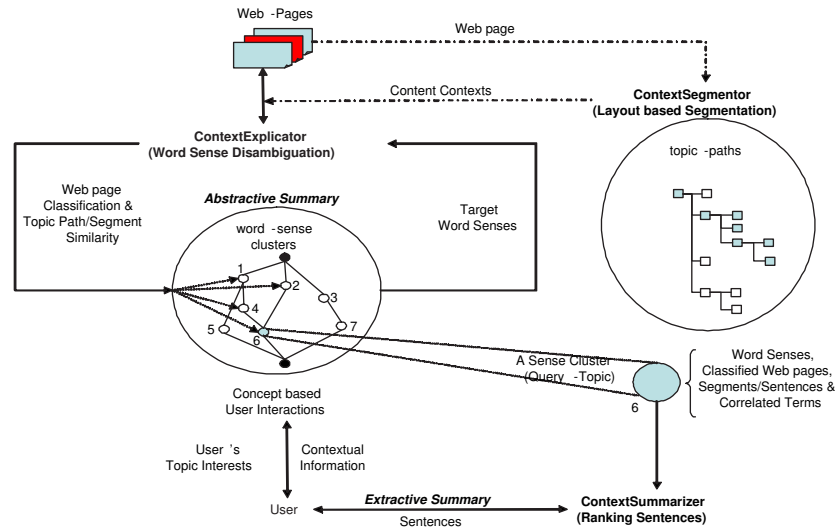
Future Web pages summarization systems might support both abstractive summary and extractive summary. An abstractive summary would provide structured meta-data (i.e., a collection of topic-index that describes itself and points out relevant Web pages) that make it possible to navigate the mass of information, within the given Web pages, at higher abstraction level. An extractive summary would show the most salient sentences within topically relevant Web pages, which could be used as the coherent information source for many applications. ContextSummarizer aims to generate an abstractive summary and extractive summaries, which are seamlessly connected each other for user-interested topics.

ContextSummarizer can be differentiated by two facts from Newsblaster: First, for topic-focused summarization, ContextSummarizer can generate extractive summaries for various query-topics (they are selected or refined by a user according to their topic interests in the given Web pages), but Newsblaster is limited to generate extractive summaries for a few predefined topics. Second ContextSummarizer provides users with a way that can simultaneously adjust

the importance weights between topic-specificity and topic-informativeness of a summary, but Newsblaster generates two different summaries that separately consider only one aspect of ‘topic-specificity’ or ‘topic-informativeness’.

Another advanced feature of our ContextSummarizer is the conceptual representation of query-topics. A concept is utilized as the basic element for the semantic communication between the contents of Web pages and a users’ topic interests about those contents. Concepts provide an efficient way minimizing Web users’ perception efforts for such semantic communication, before they lose their patience to indigestible information (e.g., huge and topically unorganized information).

### 3 System Architecture



**Fig. 1.** System Architecture: ContextSummarizer cooperates with ContextSegmentor [13] and ContextExplicator [14] to discover query-topic focused “abstractive summary” and “extractive summary” from the given Web pages

ContextSummarizer cooperates with two modules (See Figure 1) : **ContextSegmentor** [13] and **ContextExplicator** [14]. First, ContextSegmentor performs the layout-based segmentation algorithm on each Web page to discover ‘topic-paths’ (i.e, the narrative structure <sup>1</sup>; collections of hierarchically

<sup>1</sup> Narrative Structure is a structural framework that underlies the order and manner in which a story or narrative is presented to a reader or viewer [15]. We notice that the layout-based segmentation algorithm on each Web page was run as a pre-process of query-topic construction, word-sense disambiguation, and Web page(s) summarization.

connected content-segments) of respective Web page. Second, ContextExplicator performs the word-sense disambiguation algorithm on each topic-path (i.e., **content context**) of a Web page to calculate its similarity with each user-selected word sense. Third, according to a specific threshold value of the similarity measure, Web pages could be classified into several word senses. The co-occurred word senses within a set of Web pages construct a sense cluster that is contextualized by several features. The contextual features are the classified Web pages into those co-occurred word senses, the semantic matched topic-paths of the classified Web pages, and the location information (i.e., the identification numbers of Web pages, topic-paths, segments and sentences) of terms occurred within the semantic matched topic-paths. The associations among such sense clusters are represented with a conceptual lattice. In this paper, we call the conceptual lattice as the “abstractive summary” of the given Web pages. Finally, **ContextSummarizer** executes a Web pages Summarization Algorithm on a user-selected sense-cluster (e.g., node 6 in Figure 3) to mine the most important sentences <sup>2</sup> that are called as “extractive summary” in this paper. We notice that this paper is focused on discussing about the generation of an extractive summary respecting the semantic of a query-topic that is represented with a sense-cluster in an abstractive summary (See [14] for more details about the construction of an abstractive summary).

## 4 Summarization Algorithm

The importance of a semantic alignment problem between a user’s query-topic and an extractive summary was discussed in Section 2. To answer such a semantic alignment problem, the specificity value referring to a user-selected query-topic and the informativeness value representing the whole contents of relevant Web pages are considered by ContextSummarizer.

Some notations used to explain our summarization algorithm are defined as follows:

- $q_t$ : a query term;
- $W$ : the originally given Web pages  $\{w_i | i > 0\}$ ;
- $WS$ : all known word-senses  $\{\sum_t WS(q_t) | WS(q_t)$  returns the known word-senses of  $q_t\}$  for all query terms  $\{q_t | t > 0\}$ ;
- $sc_c$ : a sense cluster contextualized with a set of word-senses  $WS_c (\subset WS)$  and a set of Web pages  $W_c (\subset W)$  including contents which are sense-matched with all word-senses of  $WS_c$ ;
- $qt_k$ : a sense cluster  $sc_c$  (where  $k=c$ ) that is selected by a user as a query-topic. The Web pages  $W_k$  are used as the source for the generation of an extractive summary;
- $CE_{k**}^*$ : a content-element related to a query-topic  $qt_k$ . The subsumption relationship among content-elements, having different granularities, can

<sup>2</sup> The importance of each sentence depends on the combined importance value of the specificity value and informativeness value. See Section 4, for more details.

be illustrated as a sentence  $CE_{klmno}^{st}$  < a segment  $CE_{klmn}^{seg}$  < a topic-path  $CE_{klm}^{tp}$  < a Web page  $CE_{kl}^w$ ;  
 $CT(ws_{kq})$ : the context-words of a word sense  $ws_{kq} \in WS_k$ ;  
 $CTentW(CE_{k**}^*)$ : representative content-words occurred within a content boundary of  $CE_{k**}^*$ ;

#### 4.1 The Specificity Values of Context-Elements referring to a Query-Topic

$$S(CE_{k**}^*, qt_k) = \sum_q R(CT(ws_{kq}), CE_{k**}^*) \quad (1)$$

Equation 1 measures the specificity value, referring to the user-selected query-topic  $qt_k$ , of a context-element  $CE_{k**}^*$ . As the semantic of the query-topic  $qt_k$  can be indicated by the relevant word-senses  $WS_k$ , it calculates the sum of similarities between each word-sense  $ws_{kq} \in WS_k$  and the content-element  $CE_{k**}^*$ .  $R(CT(ws_{kq}), CE_{k**}^*)$  is calculated by the sum of  $TF * ICEF$  (it is applied from the  $TF * IDF$  in [16]) for all context-word  $ct \in CT(ws_{kq})$ ; the number  $TF$  of occurrences of a context-word  $ct \in CT(ws_{kq})$  within a content-element  $CE_{k**}^*$  is multiplied by the inverted content-element frequency  $ICEF$ , at the same granularity with the content-element  $CE_{k**}^*$ , of the context-word.

All context-elements belonging to the same granularity with  $CE_{k**}^*$  were ranked according to their specificity values referring to the user-selected query-topic  $qt_k$ , with one constraint: when a content-element has non-zero specificity values for more number of word-senses, the content-element was higher-ranked than other content-elements having non-zero specificity value for less number of word-senses.  $T$  top-ranked content-elements<sup>3</sup> became the targeted content-elements to discover their informativeness values.

#### 4.2 The Informativeness Values of Context-Elements

The  $T$  content-elements, selected according to the ranks of specificity values, might be not always matched with the  $T$  top-ranked informative (i.e., representative) content-elements of the Web pages  $W_k$ . Two possible reasons are as following:

1. During an iterative specialization procedure of a word-sense  $ws_{kq}$  relevant to a query-topic  $qt_k$ , ContextExplicator attempted to recommend more representative context-words that can distinguish relevant content-elements to the word-sense. However, our observation says that users preferred locally distinctive context-words (i.e., they were too specific context-words to distinguish all relevant content-elements to the word-sense) to globally distinctive context-words (i.e., they are reasonably general context-words to recall all relevant content-elements to the word-sense).

<sup>3</sup>  $T$  depends on the compression rate (e.g., the number of sentences in a summary versus the total number of sentences in target Web pages) requested by the user for the summary.

- ContextExplicator allows for a user to construct a new conceptual-model by conjunctively combining predefined word-senses. The problem is that the context-words defining a word-sense were usually recommended according to their occurrence-frequencies within different source data (e.g., different Web pages, topic-paths, segments or sentences) at different time. Thus, it is necessary to reconsider their informativeness over the contents of new target data source, when users want to reuse predefined word-senses and their context-words as the features for new target data source.

With such reasons, ContextSummarizer needs to consider not only the specificity values (i.e., *distinction ranks*; they reflect the relevance value of each content-element to a query-topic  $qt_k$ ) but also the informativeness values (i.e., *coverage ranks*; they reflect the coverage value of each content-element to all content-elements at the same granularity) of content-elements. ContextSummarizer measures the informativeness of a content-element with a following equation:

$$I(CE_{kp}^*) = \frac{WF(CTentW_{kp}, CE_{kp}^*)}{\sum_s WF(CTentW_{ks}, CE_{ks}^*)} \quad (2)$$

, where  $CE_{kp}^*$  is a content-element, and  $\{CE_{ks}^* | s\}$  are all content-elements (including  $CE_{kp}^*$ ) having the same granularity with  $CE_{kp}^*$ .  $I(CE_{kp}^*)$  has value-range of  $[0, 1]$ . We define the representative content-words  $CTentW_{k**}$  (here  $k** \equiv ks$  or  $kp$ ) for the content-element  $CE_{k**}^*$  as a set of unique content-words (i.e., noun or noun phrases) having equal or more occurrence-frequency ( $\neq 0$ ) than any context-word  $cw_g \in CT(ws_{kq})$  within  $CE_{k**}^*$ . Function  $WF()$  returns the occurrence frequency of the content-words  $CTentW_{k**}$  within the content of  $CE_{k**}^*$ . The informativeness (i.e., representativeness) of a content-element  $CE_{kp}^*$  is indicated by the normalized frequency within  $CTentW_{kp}$  of the content-element  $CE_{kp}^*$  over the sum frequency of  $CTentW_{ks}$  within all content-elements  $\{CE_{ks}^* | s\}$ .

### 4.3 Combining the Specificity and Informativeness of a Context-Element

ContextSummarizer uses the following equation to combine the specificity value and the informativeness value of a content-element  $CE_{kp}^*$ :

$$SI(CE_{kp}^*) = \alpha * \frac{S(CE_{kp}^*, qt_k)}{M_s} + \beta * I(CE_{kp}^*) \quad (3)$$

, where  $M_s = Max(S(CE_{k**}^*, qt_k))$  used to convert the value-range of  $S(CE_{kp}^*, qt_k)$  into  $[0,1]$ . If  $M_s = 0$  then it sets  $M_s = 1$ .  $0 \leq \alpha \leq 1$  and  $\beta = 1 - \alpha$ .  $\alpha$  and  $\beta$  is respectively the relative weight of specificity and informativeness.  $\alpha$  and  $\beta$  are adjustable according to a user's request. The default value of  $\alpha$  is 0.5. When  $\alpha > \beta$ , our system returns a more query-topic focused summary rather than a generic summary. The content-elements are ranked according to their  $SI(CE_{kp}^*)$  values.

#### 4.4 Sentences for a Web pages Summary

To extract sentences composing a summary, ContextSummarizer iteratively ranks the content-elements at different granularity from topic-path to sentence. For example,  $T_1$  top-ranked topic-paths became the source to discover  $T_2$  top-ranked segments, and the  $T_2$  top-ranked segments became the source to mine  $T_3$  top-ranked sentences. Finally, ContextSummarizer applied following four constraints (they are adjustable by the user at summarization run time) over  $T_3$  sentences to select sentences for a summary.

1. Compression Rate: the total number of selected sentences for a summary is controlled by the requested percentages of top-ranked sentences in relation to the total number of sentences of the target Web pages.
2. Length: it removed sentences that do not contain sufficient number of content-words.
3. Redundancy: if a sentence does not contain sufficient number of unique content-words against each higher ranked sentence, the sentence is removed.
4. Abstraction Level: The abstraction level of an extractive summary is indirectly adjusted according to the user-selected sense-cluster. For example, the summary for a sense-cluster (i.e., a topic) would be topically more specific compared to those of its ancestor sense-clusters (i.e., super-topics), and would be topically more general description compared to those of its descendant sense-clusters (i.e., sub-topics).

The finally selected and ranked sentences are grouped and presented to users according to the original orders in the target Web pages.

## 5 Evaluation

In [17], existing summarization evaluation techniques are subdivided into two categories: an intrinsic approach and an extrinsic approach. The intrinsic approach independently evaluates a summary from any application purposes. One way of the intrinsic approach is to comparing a summary generated by a summarization system with a human-produced “gold” summary. However, this approach has one critical difficulty of achieving people agreement on what constitutes a “gold” summary [18]. Especially, when we consider that different summaries could be “good” summaries according to a user’s different topic interests over the same documents, there would be more difficulties in getting people agreement. Thus, in this evaluation, we substituted showing a few comparable summarization results for an intrinsic approach. (We entrust the quality comparison task over those summarization results to each reader of this paper). The extrinsic approach compares the performances of an application that utilizes the system-generated summaries for its specific task. In this evaluation, we showed how the summaries extracted by ContextSummarizer can be utilized to improve clustering performances.

## 5.1 Case Studies

A major deficiency of general summarization systems, antithetical to topic-focused summarization systems, is that their summaries are significantly different at the same compression rate even over the same document(s). For example, in the “2. A major deficiency of general summarization systems” of [19], Copernic Summarizer and Pertinence Summarizer showed noteworthy differences in their summaries over the same Web page ”TWP1”. However, without any topic-focus respecting a user’s specific information needs, it is not possible to even judge which system generated better quality summaries.

Columbia Newsblaster [5] makes multi-documents, general purpose summaries on topic clusters. The topic clusters are based on six predefined categories (i.e., main topics) having hierarchical sub-topic levels. For example, in the “3.1 Multi-WebPages Summarization by Columbia Newsblaster [19]”, Columbia Newsblaster generated a summary that is talking about “festival” (subtopic) under the category of “entertainment” (topic). The Web pages clustered into the same topic cluster are assumed to provide topically coherent contents to generate a multi-documents summary. However such topical coherence might be guaranteed only in the context of predefined topics and subtopics. A problem is that any predefined topic categories cannot cover a user’s new topic interests. For example, when a user requests a summary about ”ticketing of the festival”, Columbia Newsblaster can not answer it.

ContextSummarizer resolved those problems, mentioned above, by constructing topic and subtopic categories at summarization run time by considering not only content contexts of the target Web pages but also a query-topic. For example, let us assume that a user is looking for news about “ticketing of the festival”, and another user wants to get news about “music event in the festival”. The summaries, generated by Contextsummarizer that respectively respects the users’ different interests, are displayed in the “a) and b) of 3.2 Multi-WebPages Summarization by ContextSummarizer [19]”.

## 5.2 Comparisons of Clustering Performances

In this subsection, we consider that the content-words occurred within a better summary can be used as more accurate features to properly subgroup the given test Web pages into topically relevant clusters.

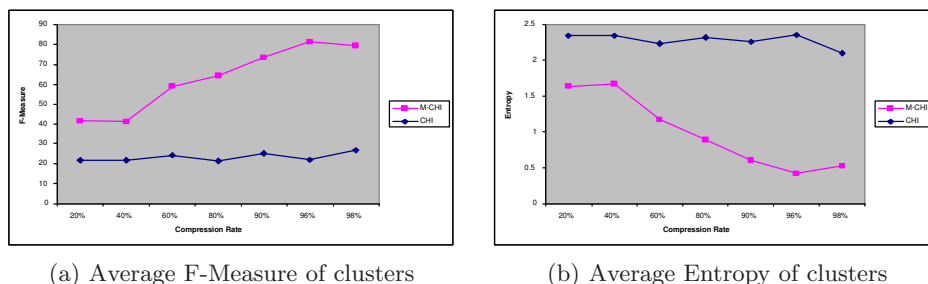
**Experiment Set Up:** The user-selections for the context-words of a query-topic might be diverse according to different users. In our experiments, we eliminated such user-dependent deviations to settle down a reasonable test circumstance for the evaluation of an extractive summary, by assuming that users’ common topic interests (i.e., query-topics) are the main topics occurred within the given test Web pages. For the test Web pages, we collected 150 documents (i.e., articles from 02/04/2006 to 02/07/2006) about 19 news topics in Columbia Newsblasters.

**Evaluation Measure:** 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 F-measure of all obtained clusters is defined by the average sum of the F-measure of each obtained cluster. The traditional F-measure (i.e., F1 measure; recall and precision are evenly weighted) of a obtained cluster is defined as  $F(W) = \frac{2 \times Precision(W) \times Recall(W)}{Precision(W) + Recall(W)}$ .

**Evaluation Results:** We compared the quality of clusters that are constructed by utilizing a set of content-words extracted with a supervised feature selection method,  $\chi^2$  statistic [20]<sup>4</sup>, from two different content sources : One content source is a set of ranked content-words  $S_1$  occurred within the given test Web pages. Another content source is a set of ranked sentences  $S_2$  for specific topic(s), by our ContextSummarizer. A compression rate is applied to two different sources  $S_1$  and  $S_2$  for the selection of the final content-words that are used as the features for the clustering of the given test Web pages. For example, to apply compression rate 5% to content source  $S_1$ , it selects the 5% top-ranked content-words  $C_1$  from  $S_1$ . To apply compression rate 5% to content source  $S_2$ , it selects top-ranked sentences  $S_2$  that contains the same number of content-words with  $C_1$ , and then calculates the weights of the content-words  $C_2$  occurred within  $S_2$ .  $C_1$  and  $C_2$  are respectively used to construct two different groups of clusters, and then the clustering performances for each group of clusters are compared.

Figure 2 shows the average F-measure and Entropy of generated clusters, at each compression rate 20, 40, 60, 80, 90, 96 and 98%. Through all compression rates, the clustering performances based on the M-CHI method, using the content-words  $C_2$  extracted from the sentences summarized by ContextSummarizer, were much better than those based on the CHI method, using the content-words  $C_1$  extracted from the test Web pages. The trend of F-measure and Entropy measure of M-CHI was getting better when the compression rate became higher (up to 96%). It can be understood as the evidence of two facts: 1) ContextSummarizer returned “good” summarized sentences that are including topically coherent information to properly cluster relevant Web pages into the tested 19 topics, and 2) A reasonable summarization performance of ContextSummarizer was kept at high compression rates. It means that much fewer

<sup>4</sup>  $\chi^2$  statistic is a supervised feature selection method which measures the statistical significance of association between a term and a category by using known class label information.



**Fig. 2.** CHI and M-CHI methods utilized the content-words, respectively extracted from the test Web pages and the summaries (i.e., sentences) generated by ContextSummarizer, to construct clusters

(i.e., human readable number of) sentences can be presented as an extractive summary to the user instead of very large Web pages.

## 6 Conclusion

ContextSummarizer is an autonomous model that adapts its summarization results according to: a) a requested query-topic (i.e., contextual features describing the user’s attention to partial contents of the given Web pages). The user can easily move their topical interests in one of the sense-clusters composing a concept-model, b) the semantic-aligned contents (i.e., Web pages, topic-paths, segments and sentences) to the query-topic, c) the redundancy removal from the topically matched and top-ranked sentences. It provides a way to compact the summary sentences extracted from large scaled Web pages, and d) the requested compression rate of an extractive summary.

Some case studies and experimental results showed that a query-topic focused extractive summary can be composed of more topically consistent sentences rather than a generic summary.

## References

1. Shen, D., Chen, Z., Yang, Q., Zeng, H.J., Zhang, B., Lu, Y., Ma, W.Y.: Web-page classification through summarization. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004, ACM (2004) 242–249
2. Amitay, E., Paris, C.: Automatically summarising web sites - is there a way around it? In: Proceedings of 9th International Conference on Information and Knowledge Management, CIKM 2000, ACM (2000)
3. : (CopernicSummarizer: <http://www.copernic.com/en/products/summarizer/>)
4. : (PertinenceSummarizer: [http://www.pertinence.net/index\\_en.html](http://www.pertinence.net/index_en.html))
5. : (Columbia Newsblaster: <http://www1.cs.columbia.edu/nlp/newsblaster/>)

6. Alam, H., Hartono, R., Kumar, A., Rahman, F., Tarnikova, Y., Wilcox, C.: Web page summarization for handheld devices: A natural language approach. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition Volume II, ACM (2003) 1153–
7. Berger, A.L., Mittal, V.O.: *ocelot*: a system for summarizing web pages. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2000, ACM (2000) 144–151
8. Buyukkokten, O., Garcia-Molina, H., Paepcke, A.: Seeing the whole in parts: text summarization for web browsing on handheld devices. In: Proceedings of the Tenth international conference on World Wide Web, WWW 2001, ACM (2001) 652–662
9. Delort, J.Y., Bouchon-Meunier, B., Rifqi, M.: Links for a better web: Enhanced web document summarization using hyperlinks. In: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia, ACM (2003) 208–215
10. Delort, J.Y., Bouchon-Meunier, B., Rifqi, M.: Web document summarization by context. In: Proceedings of the 12th international conference on World Wide Web - Alternate Track Papers & Posters, WWW 2003, ACM (2003)
11. Jatowt, A.: Web page summarization using dynamic content. In: Proceedings of the 13th international conference on World Wide Web - Alternate Track Papers & Posters, WWW 2004, ACM (2004) 344–345
12. Jatowt, A., Ishizuka, M.: Temporal web page summarization. In: Proceedings of 5th International Conference on Web Information Systems Engineering, WISE 2004, Springer (2004) 303–312
13. Yoo, S.Y., Hoffmann, A.: Pseudo-relevance feedback in web information retrieval using segments' subjective importance values. In: 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), IEEE Computer Society (2005) 422–425
14. Yoo, S.Y., Hoffmann, A.: Knowledge-level management of web information. In: 6th Asia-Pacific Web Conference, APWeb 2005, Springer (2005) 597–608
15. : Wikipedia (2006) [http://en.wikipedia.org/wiki/Narrative\\_structure](http://en.wikipedia.org/wiki/Narrative_structure).
16. Salton, G., Buckley, C.: Term-weighting approaches in automatic retrieval. *Information Processing and Management* **24** (1988) 513–523
17. Mani, I., Maybury, M.T.: *Advances in Automatic Text Summarization*. MIT Press (1999)
18. Stergos D. Afantenos, V.K., Stamatopoulos, P.: Summarization from medical documents: a survey. *Artificial Intelligence in Medicine* **33** (2005) 157–177
19. : ContextSummarizer: <http://www.cse.unsw.edu.au/~syy/ContextSummarizer-casestudy.html>. (2005)
20. 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