Hierarchical Soft Clustering and Automatic Text Summarization for Accessing the Web on Mobile Devices for Visually Impaired People

Gaël Dias, Sebastião Pais, Fernando Cunha, Hugo Costa, David Machado, Tiago Barbosa and Bruno Martins

Centre of Human Language Technology and Bioinformatics, University of Beira Interior, Covilhã, Portugal
{ddg, sebastiao, fernando, hugo, david, tiago, bruno}@hultig.di.ubi.pt

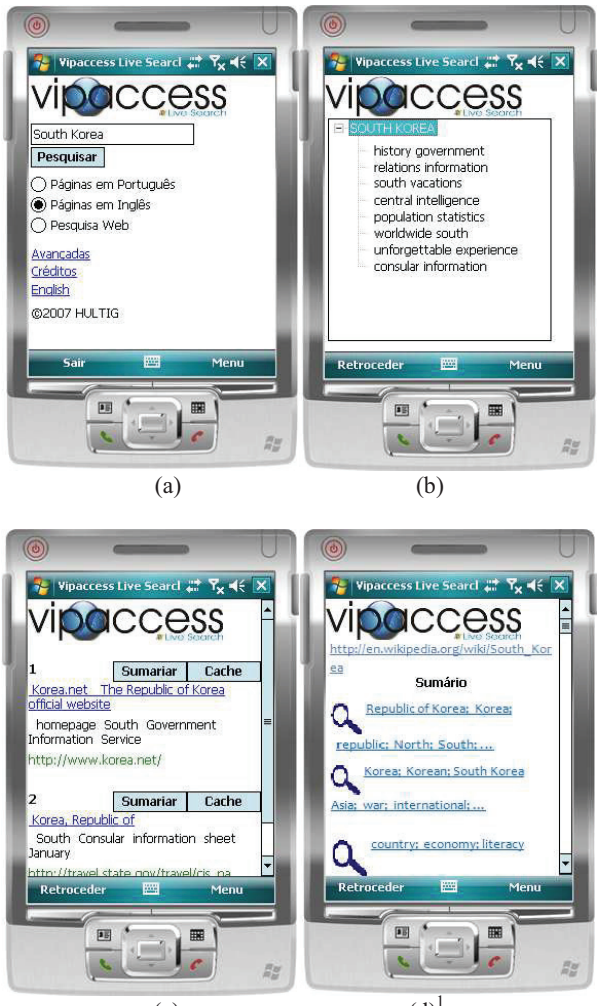
Abstract

In this paper, we propose a universal solution to web search and web browsing on handheld devices for visually impaired people. For this purpose, we propose (1) to automatically cluster web page results and (2) to summarize all the information in web pages so that speech-to-speech interaction is used efficiently to access information.

Introduction

Although a lot has been done for Visually Impaired People (VIP) to access information with Braille screens, Braille keyboards, Braille handheld devices and speech-to-speech interfaces, very little has been done to reduce the amount of information they have to deal with. Indeed, blind people face an overwhelming task when reading texts. Unlike fully capacitated people, blind people cannot read texts by just scanning them quickly i.e. they cannot read texts transversally. As a consequence, they have to come through all the sentences of a text to understand its level of importance. In the specific context of Information Retrieval, two issues must be particularly tackled for universal access to information: web search and web browsing. In the first case, web search engines usually provide long lists of unorganized search results in the form of web page titles and corresponding snippets. This way of presenting information is a clear obstacle for VIP for quick access to information as long lists of results take long time to scan. In the second case, web pages are usually designed to present information in a unique form independently of the intended user. As a consequence, VIP find it hard to evaluate the importance of a given document as they have to come through all the sentences of the web page.

In parallel, the shift in human computer interaction from desktop computing to mobile interaction highly influences the needs for future user-adaptive systems. Indeed, small size screens of handheld devices are a clear limitation to display long lists of relevant documents resulting in time consuming scrolling. In the context of web browsing, this issue is also crucial as most web pages are designed to be viewed on desktop displays.



(a) (b)

(c) (d)[1]

Figure 1: VipAccess Meta Search Engine

[1] At the moment of writing the paper, this feature is implemented as a stand-alone module and is currently being integrated in the global architecture.

As a consequence, the smallest web page excerpts displayed on any mobile device screen can interfere with users' comprehension as repetitive zooming and scrolling may be necessary.

To overcome these issues, we propose a universal solution to mobile information retrieval. For this purpose, we propose (1) to automatically cluster web page results and (2) to summarize all the information in web pages, so that all cognitive processes of the information search are fully accessible to VIP and also fit into the small displays of mobile devices thus reducing scrolling and zooming.

To illustrate the overall process, we present the different steps performed to present web contents in Figure 1. In particular, Figure 1a displays the main search interface, Figure 1b shows the clusters automatically extracted from the web page results, Figure 1c presents all the web pages within a given cluster with their snippets summarized by a list of five relevant terms and finally Figure 1d illustrates the summary of a selected web page.

To achieve the hierarchical structure, we implement a soft clustering algorithm which allows web pages to be present in different clusters based on two representations of web snippets: language-dependent and language-independent. Then, text mining techniques are applied to summarize web snippets and web pages based on linguistic and statistical processing depending on whether it is possible to automatically identify the language of web pages or not.

Related Work

As far as we know, similar architecture has never been proposed so far. However, many works have been proposed in the context of web page results clustering and text summarization for mobile devices.

Clustering of Web Page Results

Many different works have been proposed in the area of web page results clustering based on snippets analysis. In order to cluster web page results, it is necessary to evaluate web snippet similarity. Different methodologies have been proposed. For example, (Ferragina and Gulli, 2003; Fung *et al.*, 2003) rely on choosing the *k* most significant words of the web page snippet to represent it based on the tf.idf weighting scheme. The clustering phase is then performed over a similarity matrix calculated on the vector space model using the cosine similarity measure. Others, like, (Zhang and Dong; Jiang *et al.*, 2002; Zheng *at al.*, 2004) use shared n-grams between the web snippets to compute snippet similarity.

The different approaches also distinguish themselves by the match of the following two items: (1) simple words or phrases and (2) flat or hierarchical clustering. As (Ferragina and Gulli, 2003) refer to, the simplest case is the match between flat clustering and single words, whilst the general case is the match between hierarchical clustering and phrases. In the simplest case, (Jiang *et al.*, 2001) cluster the results returned from a search engine based on a relational fuzzy clustering algorithm RFCMdd, based on the idea of identifying k-medoids. (Zeng *et al.*, 2004) extract phrases as candidate cluster names based on a regression model over five different characteristics. A phrase has a document associated to it and the clusters are generated by merging the documents that share the same phrases which exceed a certain threshold. In the context of hierarchical clustering, (Fung *et al.*, 2003) propose hierarchical clustering with single words. The clustering process is based on the idea of frequent itemsets which are a set of words that occur together in some minimum fraction of web pages in a given cluster. (Ferragina and Gulli, 2003) use a snippet analyzer extracting phrases such as named entities. The hierarchy construction process consists in grouping documents sharing the same sentence. Finally, (Zhang and Dong; 2001) use an algorithm based on suffix arrays for keyphrase discovery and SVD to reduce data sparseness before hierarchical clustering.

Text Summarization for Mobile Devices

Many different works have been proposed in the area of text summarization. However, we will exclusively focus on the methodologies applied to handheld devices. (Buyukkokten et al., 2000) is certainly the most relevant first appearing paper of this field introducing two methods for summarizing parts of web pages. Each web page is broken into Semantic Textual Units (STUs) that can each be hidden, partially displayed, made fully visible, or summarized. However, their work is built on old well known techniques for text summarization and do not introduce linguistic processing (except stemming) to remain real-time adaptable as processing is handled by the mobile device. In order to introduce more knowledge compared to the previous model, (Yang and Wang, 2003) propose a fractal summarization model based on statistical and structure analysis of web pages. Thus, thematic features, location features, heading features, and cue features are adopted. Their architecture first generates a skeleton of a summary and its details are generated on user request. Comparatively to (Buyukkokten et al., 2000), they propose a more organized structure but do not use any linguistic processing although they work on basis of a three-tier architecture which provides more processing power. (Gomes et al., 2001) are the first to introduce some linguistic knowledge into the process of text summarization. They use a parser to perform text segmentation and morphological analysis. In particular, they apply linguistic patterns for sentence compression rather than for sentence extraction. For example, some names are replaced with their acronyms and some adjectives may also be removed. The major drawback of this approach is the lack of statistical analysis which is a key factor for high quality summarization.

Our Contribution

In this paper, we implement hierarchical clustering with phrases. For this purpose, we use a soft clustering algorithm called PoBOC (Cleuziou *et al.*, 2004) and a statistical phrase extractor called SENTA (Dias *et al.*, 1999) which are both parameter free, thus allowing their application to real-world environments, contrarily to most of the methodologies proposed so far in the literature.
In the field of text summarization, we propose to use both statistical evidence and linguistic processing for sentence extraction in real-time. For this purpose, we propose a new sentence weighting scheme based on the TextRank algorithm (Mihalcea and Tarau, 2004).

Hierarchical Clustering

The hierarchical structure of web page results is obtained by the soft clustering algorithm PoBOC (Cleuziou *et al.*, 2004) applied over two web snippet representations.

Web Page Representation

Unlike all other methodologies which treat all query results in the same way, we propose to take advantage of linguistic processing when language identification is possible. So, if all web page results are in English or Portuguese, the snippets are linguistically processed before clustering is performed. Otherwise, web snippets are treated as bags of relevant words and phrases[2].

Language Identification

To identify the language of a text such as proposed in (Beesley, 1998), a distance between its frequency-ordered list of character n-grams and language baseline frequency ordered-lists can be computed. For each n-gram in the test document, there can be a corresponding one in the current language profile it is compared to. N-grams having the same rank in both profiles receive a zero distance. If the respective ranks for an n-gram vary, they are assigned the number of ranks between the two. Finally all individual n-gram rank distances are added up and evaluate the distance between the sample document and the current language profile. In our context, a text is a snippet which is compared to two reference frequency lists of 2-grams: one for English and one for Portuguese.

Language-Dependent Representation

If it is possible to identify the language of a web snippet, it is morpho-syntactically tagged using the TreeTagger (Schmid, 1994) in order to process clustering on more reliable basis. The statistical phrase extractor SENTA (Dias *et al.*, 1999) is then applied to identify relevant phrases from all the web page results. Finally, only the lexical items that respect the following regular expressions (respectively for English and Portuguese) are kept to represent the contents of web snippets:

[Noun | Verb | Adj | Noun Noun$^+$ | Adj Noun$^+$ | Noun Prep Noun],
[Noun | Verb | Adj | Noun Noun$^+$ | Noun$^+$ Adj | Noun Prep Noun].

In order to perform clustering, each feature of the web snippet (i.e. nouns, verbs, adjectives and compound nouns) must be given a relevance value. For this purpose, we propose a new weighting scheme based on frequency and position of appearance. In the context of meta-search engines, the relative frequency of a single word w can be expressed as in Equation (1) where $|Snippets|$ represents the number of web snippets retrieved for each url and $f(y)$ is the number of occurrence of any word y.

$$\hat{f}(w) = \frac{\sum_{i=1}^{|Snippets|} f(w)}{\sum_{i=1}^{|Snippets|} \sum_{y \in snippet\{i\}} f(y)} \quad (1)$$

The exact same process is then applied to evaluate the relative frequency of any compound noun cn as shown in Equation (2) where $f(cnj)$ is the number of occurrence of any compound noun cnj.

$$\hat{f}(cn) = \frac{\sum_{i=1}^{|Snippets|} f(cn)}{\sum_{i=1}^{|Snippets|} \sum_{cnj \in snippet\{i\}} f(cnj)} \quad (2)$$

It has also been proven that word position in texts plays an important role for the evaluation of its importance (Witten *et al.*, 1999). As a consequence, we propose the following weighting factor $\hat{o}(.)$ which increases the importance of words or compound nouns w appearing in the beginning of web snippets.

$$\hat{o}(w) = \frac{-\sum_{i=1}^{|Snippets|} \ln \frac{occur(1, w, snippet\{i\})}{card(snippet\{i\})}}{|Snippets|} \quad (3)$$

In Equation (3), $occur(1,w,snippet\{i\})$ is the function that retrieves the first position of the word w in the snippet $snippet\{i\}$ and $card(snippet\{i\})$ stands for the number of words in the snippet $snippet\{i\}$[3]. So, the final weight W of any feature of the web snippet (i.e. noun, verb, adjective or compound noun) is evaluated as in Equation (4) and (5) whether it is a single word or a compound noun. Indeed, as compound nouns convey valuable information which cannot be compressed into a single word, an extra weight is given by summing all the relative frequencies of its

[2] This process is fundamental for the selection of the corresponding speech engine.

[3] The first occurrence of a compound noun is the occurrence of its first occurring word.

constituents. The ||.|| operator means that the values have been normalized so that all attributes can be compared on the same scale.

$$W(w) = \left\| \hat{f}(w) \right\| + \left\| \hat{o}(w) \right\| \quad (4)$$

$$W(cn) = \left\| \hat{f}(cn) \right\| + \left\| \hat{o}(cn) \right\| + \sum_{w=1}^{|cn|} \left\| \hat{f}(w) \right\| \quad (5)$$

Language-Independent Representation

If it is impossible to identify the language of a web snippet, it is represented as a bag of relevant words. For this purpose, the web snippets are tokenized and SENTA is applied to all retrieved web snippets to identify statistically relevant phrases from all the web page results. As a consequence, each web snippet is represented by its list of words in which relevant phrases have been identified.

Web Page Clustering

In order to cluster web page results, we propose to use the PoBOC algorithm (Pole-Based Overlapping Clustering). A recent comparative study by (Cicurel et al., 2006) shows that CBC (Clustering By Committees) and PoBOC both lead to relevant results for the task of word clustering. Nevertheless CBC requires parameters hard to tune whereas PoBOC is free of any parameterization. Moreover, unlike most of commonly used clustering algorithms, PoBOC shows the following advantages among others: (1) the input is restricted to a single similarity matrix, (2) the number of final clusters is automatically found and (3) it provides overlapping clusters allowing taking into account the different possible meanings of web page results.

Similarity Matrix

As a consequence, we propose to build two square similarity matrixes where each cell corresponds to the similarity between two web page results depending on whether language identification is possible or not. For the language-dependent representation, each url is compared to any of the other urls based on the classical cosine similarity measure over the attribute-value representation of web snippets presented above (See Equation 6).

$$\cos(url1, url2) = \frac{\sum_{w} W_{url1}(w) \times W_{url2}(w)}{\sqrt{\sum_{w} W_{url1}(w)^2} \times \sqrt{\sum_{w} W_{url2}(w)^2}} \quad (6)$$

For the language-independent representation, we propose to measure the similarity between two urls by applying the string distance Sumo-Metric (Cordeiro et al., 2007) to their web snippets. In particular, the Sumo-Metric has proved to lead to improved results when compared to Word N-gram Overlap techniques. It is defined in Equation (7) where $|url|$ stands for the number of words in the web snippets representing the *url*, λ the number of exclusive links between common words in both web snippets, and α and k are tuning parameters[4].

$$S(url1, url2) = \begin{cases} S(|url1|, |url2|, \lambda), \text{if } S(|url1|, |url2|, \lambda) < 1 \\ e^{-k * S(|url1|, |url2|, \lambda)}, \text{otherwise} \end{cases}$$
$$\text{where}$$
$$S(|url1|, |url2|, \lambda) = \alpha \log_2 \frac{|url1|}{\lambda} + (1-\alpha) \log_2 \frac{|url2|}{\lambda} \quad (7)$$

Hierarchical Soft Clustering

Based on the similarity matrix, the PoBOC algorithm first builds a similarity graph and evaluates local maxima of similarity values. Then, it searches for poles in the similarity graph i.e. clique sub-graphs. Then, it evaluates the membership of each url to all the poles and assigns the url to one or several poles. A final step consists in organizing the obtained clusters into a hierarchical tree. This process is accomplished by recursively applying the PoBOC algorithm to the cluster similarity matrix where the similarity between clusters is computed based on their centroids in the same way it is done for the urls.

Differential Cluster Labeling

Human users interact with clusters. As such, clusters need to be labeled so that users can see what a cluster is about. For this purpose, we implement the well-known differential cluster labeling technique (Manning *et al.*, 2008) which selects cluster labels by comparing the distribution of terms in one cluster with that of other clusters. As a consequence, for the language-dependent representation, the candidate for cluster label will be the noun or compound noun which (1) occurs in most of the urls of the cluster and (2) evidences high relevance to the cluster when compared to the other ones as in Equation (8) where $|cluster|$ stands for the number of urls in the cluster, $|cluster, w|$ the number of urls containing the word w, N the number of clusters and Nw the number of clusters containing w.

$$r(w, cluster) = \frac{|cluster, w|}{|cluster|} \times \log_2 \frac{N}{Nw} \quad (8)$$

Similarly, for the language-independent representation, the candidate for cluster label is the word which evidences a maximum $r(.,.)$ value.

Text Summarization

In the first part of this paper, we proposed to automatically cluster web pages results to reduce scrolling and zooming for small handheld devices as shown in Figure 1b. However, our task does not end here. Scrolling and zooming must also be avoided for web page contents. For this purpose, we propose a new architecture for

[4] α and k are respectively set to 0.5 and 3 in our architecture.

summarizing Semantic Textual Units (Buyukkokten et al., 2000) based on efficient algorithms for linguistic treatment (Gil and Dias; 2003; Schmid, 1994) that allow real-time processing and deeper linguistic analysis of web pages to produce quality content visualization.

Semantic Textual Units

One main problem to tackle is to define what to consider as a relevant text in a web page. Indeed, web pages often do not contain a coherent narrative structure (Manning et al., 2008). So, the first step of any system is to identify rules for determining which text should be considered for summarization and which should be discarded. For this purpose, (Buyukkokten et al., 2000) propose Semantic Textual Unit (STU) identification. STUs are page fragments marked with HTML markups which specifically identify pieces of text following the W3 consortium specifications. It is clear that the STU methodology is not as reliable as any language model for content detection but on the opposite it allows fast processing of web pages.

Linguistic Treatment

On the one hand, single nouns usually convey most of the information in written texts. On the other hand, compound nouns are frequently used in everyday language, usually to precisely express ideas and concepts that cannot be compressed into a single word. As such, each STU in the web page is part-of-speech tagged with the TreeTagger and processed with SENTA to extract relevant phrases in it. Some well-known heuristics are then applied to define quality compound nouns. So, only multiword units that respect the following regular expressions (respectively for English and Portuguese) are tagged as single words:

[Noun Noun$^+$ | Adj Noun$^+$ | Noun Prep Noun],
[Noun Noun$^+$ | Noun$^+$ Adj | Noun Prep Noun].

In the case language identification is impossible, the STUs are simply tokenized and processed with SENTA to identify and mark statistically relevant phrases.

Extractive Text Summarization

Extractive text summarization aims at finding the most significant sentences in a given text. So, a significance score must be assigned to each sentence in a STU. The sentences with higher significance naturally become the summary candidates and a compression rate defines the number of sentences to extract. For this purpose, we implement the TextRank algorithm (Mihalecea and Tarau, 2004) combined with an adaptation of the well-known inverse document frequency, the inverse STU frequency (*isf*) to weight word relevance. The basic idea is that highly ranked words with high *isf* are more likely to represent relevant words in the text and as a consequence provide good clues to extract relevant sentences for the summary.

In our case, each STU is first represented as an un-weighted oriented graph being each word connected to its successor following sequential order in the text as shown in Figure 2 for the following sentence: "*The British Council of Disabled People is the UK's National Organization of the worldwide Disabled People's Movement*".

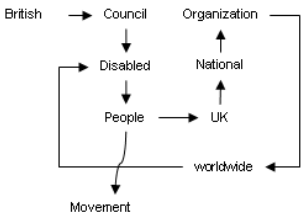


Figure 2: Graph Representation of STU.

Following the TextRank algorithm, the score of a word w_i is defined as in Equation (9) where $In(w_i)$ is the set of words that point to it, $Out(w_j)$ is the set of words that the word w_j points to and d is the damping factor set to 0.85.

$$S(w_i, stu) = (1-d) + d \times \sum_{j \in In(w_i)} \frac{S(w_j, stu)}{|\,Out(w_j)\,|} \quad (9)$$

Then, each word is weighted as in Equation (10) based on its graph-based ranking and its relevance in the text based on its inverse STU frequency (*isf*) where N is the number of STUs in the text and $stuf(w)$ is the number of STUs in which the word w appears.

$$rw.isf(w, stu) = S(w, stu) \times \log_2 \frac{N}{stuf(w)} \quad (10)$$

Finally, the sentence significance weight is defined as in (Vechtomova and Karamuftuoglu, 2004), thus giving more weight to longer sentences, in Equation (11) where |S| stands for the number of words in S, w_i is a word in S and where argmax(|S|) is the length of the longest sentence in the STU.

$$weight(S, stu) = \frac{\sum_{i=1}^{|S|} rw.isf(w_i, stu) \times |\,S\,|}{\arg\max_S(|\,S\,|)} \quad (11)$$

Visualization

The last part of the process is the visualization phase. Following the same strategy as in (Buyukkokten et al., 2000), the user is presented the five most relevant keywords of each STU as well as the importance of the STU in the overall text through a magnifying glass. By selecting the interesting STU, the user is presented the most relevant summary that fits the handheld device display i.e. the most relevant sentences computed as in Equation (11). In particular, the significance factor of a

STU (*sfactor*) is simply calculated as in Equation 12, where |*stu*| stands for the number of sentences in the STU, S_i is a sentence in the given STU and argmax(|*stu*|) is the length of the longest STU in the text.

$$sfactor(stu) = \frac{\sum_{i=1}^{|S|} weight(S_i, stu) \times |stu|}{\arg\max_{stu}(|stu|)} \quad (12)$$

This weight is then normalized among all STUs in the web page so that its value ranges between [0..1] and the magnifying glass process is eased.

Conclusions

In this paper, we propose a universal solution to mobile Information Retrieval which automatically clusters web pages results and summarizes all the information in web pages to fit into the small displays of mobile devices. For this purpose, we implement a soft clustering algorithm which allows web pages to be present in different clusters through two distinct web snippets representations: language-dependent and language-independent. As all the cognitive processes of information search are summarized, our architecture allows the access to information to Visually Impaired People. For this purpose, we implemented a speech-to-speech interface based on the Microsoft SAPI module which allows speech-to-speech interaction. On the one hand, when language identification is possible, the correct speech engine is loaded which provides the user with a text-to-speech interface. On the other hand, the user can choose its language between Portuguese and English for the speech recognition task. To assess the benefits of such architecture, an evaluation has been conducted in the context of blind users which has received very positive feedback and makes us believe that the shift in human computer interaction from desktop computing to mobile interaction is also opened to VIP.

References

Beesley, K.B.. 1998. *Language Identifier: a Computer Program for Automatic Natural-Language Identification of On-line Text*. 29th Annual Conference of the American Translators Association.

Buyukkokten, O., Garcia-Molina H., Paepcke, A.. 2000. *Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices*. 10th International World Wide Web Conference.

Cicurel, L., Bloehdorn, S., Cimiano, P.. 2006. *Clustering of Polysemic Words*. In Advances in Data Analysis. 30th Annual Conference of the German Classification Society.

Cleuziou, G., Martin, L., Vrain, C.. 2004. *PoBOC: an Overlapping Clustering Algorithm: Application to Rule-Based Classication and Textual Data*. 16th European Conference of the Association for Computational Linguistics.

Cordeiro, J.P., Dias, G., Brazdil, P.. 2007. *Learning Paraphrases from WNS Corpora*. 20th International FLAIRS Conference. AAAI Press. Key West, USA.

Dias, G., Guilloré, S., Lopes, J.G.P.. 1999. *Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora*. 6ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles. Cargèse, France.

Ferragina, P., Gulli, A.. 2003. *A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering*. 14th International Conference on Data Mining.

Fung, B., Wang, K., Ester, M.. 2003. *Large Hierarchical Document Clustering using Frequent Itemsets*. SIAM International Conference on Data Mining.

Gomes, P., Tostão, S., Gonçalves, D., Jorge, J.. 2001. *Web-Clipping: Compression Heuristics for Displaying Text on a PDA*. 3rd Workshop on Human Computer Interaction with Mobile Devices.

Jiang, Z., Joshi, A., Krishnapuram, R., Yi, L.. 2002. *Retriever Improving Web Search Engine Results using Clustering*. In Managing Business with Electronic Commerce.

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

Mihalcea, R., Tarau, P.. 2004. *TextRank: Bringing Order into Texts*. Conference on Empirical Methods in Natural Language Processing.

Gil, A., Dias, G.. 2003. *Using Masks, Suffix Array-based Data Structures and Multidimensional Arrays to Compute Positional Ngram Statistics from Corpora*. Workshop on Multiword Expressions of the 41st International Conference of the Association for Computational Linguistics.

Schmid, H.: Probabilistic Part--of--Speech Tagging Using Decision Trees. Conference on New Methods in Language Processing. (1994).

Vechtomova, O., Karamuftuoglu, M. 2004. Comparison of Two Interactive Search Refinement Techniques. Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting.

Witten, I. H., Gordon, W.P., Eibe, F., Gutwin, C., Nevill-Manning, C.G.. 1999. *KEA: Practical Automatic Keyphrase Extraction*. 4th ACM Conference on Digital Libraries.

Yang, C., Wang, F.L.. 2003. *Fractal Summarization for Mobile Devices to Access Large Documents on the Web*. International World Wide Web Conference.

Zeng, H., He, Q., Chen, Z., Ma, W.: Learning to Cluster Web Search Results. 27th Annual International Conference on Research and Development in Information Retrieval, Sheffield, UK, 210-217, (2004).

Zhang, D., Dong, Y.: Semantic, Hierarchical, Online Clustering of Web Search Results. 6th Asia Pacific Web Conference. (2001).