International Journal of Computational Intelligence Systems
Vol. **12(1)**; 2019, *pp.* 299–310
DOI: https://doi.org/10.2991/ijcis.2019.125905647; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

A Methodology to Refine Labels in Web Search Results Clustering

Zaher Salah[1,*], Ahmad Aloqaily[1], Malak Al-Hassan[2], Abdel-Rahman Al-Ghuwairi[1]

[1] *Prince Al Hussein Bin Abdullah II Faculty for Information Technology, Hashemite University, Zarqa, Jordan*
[2] *Department of Business of Information Technology, University of Jordan, Amman, Jordan*

ARTICLE INFO

Article History
Received 25 Jun 2018
Revised 19 Nov 2018
Accepted 14 Dec 2018

Keywords
Information retrieval
Machine learning
Web search results clustering
Web intelligence

ABSTRACT

Information retrieval systems like web search engines can be used to meet the user's information needs by searching and retrieving the relevant documents that match the user's query. Firstly, the query is inputted to the web search engine and assumed to be a good representative for the user's intention and reflecting specifically his information needs and thus it should be long enough, discriminative, specific and unambiguous. Secondly, the web search engine typically respond to the query by sending back a long flat list of web search results and each search result represents a relevant document. Typically, that list may contain thousands or millions of web search results and thus it is difficult to navigate and locate a specific document relevant to a specific topic. As a postretrieval process, web search results clustering may be a solution for this issue where web search results can be categorized as clusters. These clusters supposed to contain topically related documents and labelled by descriptive and concise labels. These labels supposed to correctly describe the contents of each cluster. Thus the users can easily choose a cluster representing the intended topic and navigate through relatively few documents inside that cluster. High-quality labelling for clusters is crucial for users who can now gain insight into that clusters' contents, general structure, and distribution of the topics among documents in the clusters. This make the user able to preview and navigate easily and fast. To this end, the authors in this paper introduced a methodology to enhance labels for clusters of web search results. The proposed methodology is founded on the idea of using the existing labels nominated by the original Suffix Tree Clustering (STC) algorithm and adapting these labels and/or clusters so that it become more concise and descriptive. The propose methodology was conducted on the original STC algorithm to produce an enhanced version of the classical STC algorithm. The enhanced algorithm was experimented and the produced clusters and labels were evaluated and compared with respect to the classical STC algorithm. For evaluation, the authors used clusters labelling performance measure considered five parameters f1: Comprehensibility, f2: Descriptiveness, f3: Discriminative Power, f4: Uniqueness, and f5: Nonredundancy. The reported results shown that the new enhanced labels outperformed the original labels and the overall performance has been enhanced. The recorded results indicated that: (i) The proposed methodology achieved better performance and the overall average recorded values for the used performance measure (f6) was 0.921. (ii) Number of clusters was decreased from 15 to 9 clusters only. (iii) Number of duplicated results was decreased from 143 to 121 only, and (iv) average number of phrases per label was increased from 1.67 to 2.00 phrases.

1. INTRODUCTION

Web pages are dynamic, superabundant, and miscellaneous and thus a heterogeneous variation of topics represented by this massive amount of documents is expected as these documents are originated from various resources worldwide and covering various topics like, for example, arts, science, engineering, economy, politics, sport, and so on.

Starting, typically, from the first web search result(s) the user may waste a valuable time to browse many search results which are irrelevant to the intended topic that the user looking for. As a result, nearly (50%) of users only browse the first two pages of the Web Search Results returned from the search engine [1]. This arises the need for more sophisticated web search engines that may employ "postretrieval" process to improve the presentation of the search results to the users [2] where the search results can be organized into labelled clusters. Each cluster represents a specific topic and this is very important to the user who will be able to avoid previewing many irrelevant documents by deepen the navigation inside the intended cluster covering a specific topic and containing homogenous documents relevant to the user's information needs. Furthermore, it is beneficial for the naive users to possibly find "unexpected relevant documents" too [3]. To achieve this we need more than document grouping; the most important requirement is how to choose a comprehensible and expressive descriptor or label for each documents group, so that the user will be able to locate the intended documents easier and faster depending on that short and meaningful labels to concisely explain to the users what the group's content is about [4].

In general, the performance of many search results clustering techniques is still poor in the context of clusters labelling especially when the labelling phase is highly dependent on the efficiency of

* *Corresponding author. Email:* zahersalah@hotmail.com

clustering phase that determine the content of each cluster. Homogeneous content for a cluster is crucial to induce a good label for that cluster. Many promising directions for various research approaches presented themselves so as to extend the functionality and enhance the operation of specific phases in the clustering algorithms. Choosing the appropriate clustering algorithm with optimized parameters and efficient mechanism to elect the representative terms to be used as candidate labels are the key to produce better clusters with concise and knowledgeable (descriptive and informative) labels.

Web search engines (like, for example, Google, Bing, Dogpile, Yahoo, and Baidu) respond to the user query by sending a long list of search results meeting the information requirements (user intention) expressed by that query. In ranked retrieval systems, the search results list is ordered decreasingly according to a specific relevancy ranking (scoring) scheme and typically contains a title, a small portion of text called snippet and a URL for each search result [3]. Figure 1 shows an example of web search results for the query: *Hashemite university* represented as a flat ordered list of results. Both the low precision and flat presentation of search results made the process of meeting the user's information needs far more exhaustive than it should to be and thus raise the need for more sophisticated search engines in which the relevant search results are easier to brows and to navigate [5].

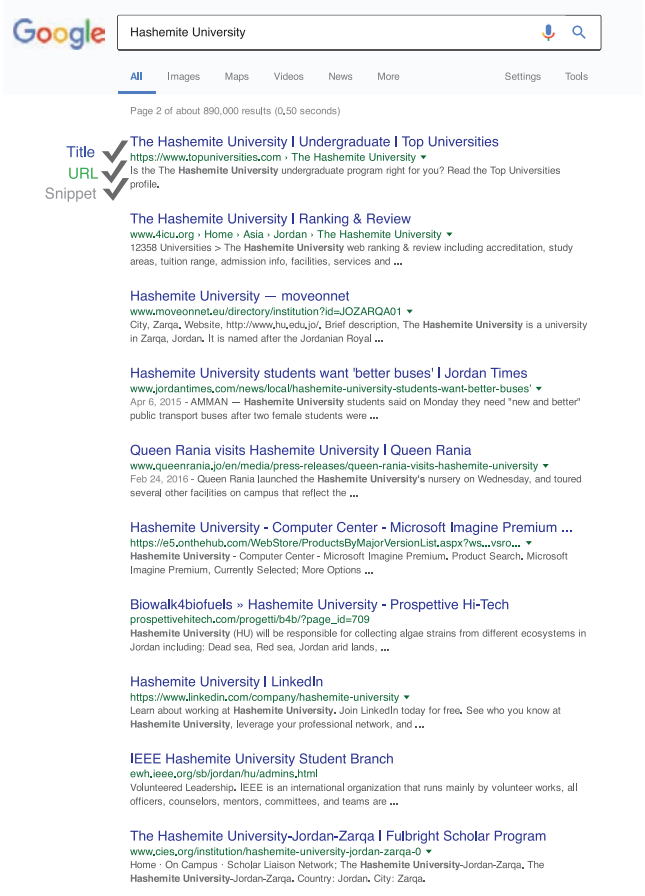

Figure 1 An example of a flat list of web search results for the query: *Hashemite university* containing a title, a URL and a snippet for each search result associated with a relevant document.

2. PRELIMINARIES

In this previous work section we provide some background concerning the labelling phase in the process of web search results clustering (WSRC). Various types of algorithms can be adopted for clustering textual documents employing, for example: neural network, fuzzy logic, rough set, or graph theory. WSRC algorithms can be classified into two categories [6]: (i) Numerical-based algorithms which have performance issues with web search result clustering because it expect full-text as input not short-text snippets like in web search results. In the other hand, its output is "raw numerical" thus it cannot be used to label clusters because it is un-interpretable by the user. (ii) Phrase-based algorithms which produce more comprehensible and descriptive labels than numerical algorithms. Table 1 [6] below differentiates between numerical-based and phrase-based algorithms while Table 2 [7] presents a comparison between the most typically used clustering algorithms in the literature based on various characteristics.

Table 1 Comparison between numerical-based and phrase-based algorithms.

Numerical-Based Algorithms	• Documents are converted to term- document matrix. • Numerical algorithms require more data than is available. • Raw numerical outcome is also difficult to convert back to cluster description • Data model is usually used Vector Space Model.
Phrase-Based Algorithms	• Phrase based on frequent phrases instead of numerical • It is simpler than numerical algorithms. • These algorithms usually discard smaller clusters. • Data model is usually used N-gram, Suffix Tree.

The Commonly used Suffix Tree Clustering (STC) algorithm deals with each document as a string (sequence of words) instead of a bag-of-words (BOW) which neglects the order of words while considers only the frequencies of distinct words occurrences in the corpus. STC uses suffix trees for summarizing the documents and extracting the frequent phrases while other algorithms like semantic hierarchical online clustering (SHOC) uses suffix arrays instead [8]. Label Induction Grouping Algorithm (Lingo) produces more clusters than STC and K-mean algorithms while STC is more scalable than Lingo and K-mean. [9] Lingo commences with extracting expressive labels first and then clustering documents individually to the fittest label (each label representing a cluster). Labels are generated from the pruned frequent terms (phrases and words) that achieve the required level of labelling descriptiveness and informativeness quality [7, 10].

In both Lingo and SHOC algorithms, labels of the clusters should be (i) present in a web search result snippet a number of times exceed a given threshold, (ii) meaningful and contained in a single sentence covering a specific topic, and (iii) clear by being complete (not partial phrase), long enough and frequent phrase. In addition, stop words that are present in the phrase must be preserved to produce more eligible cluster labels [11].

Table 2 A comparison between the most typically used clustering approaches.

Method	Semantic Relation	Cluster Label	Phrase Based	Incremental	Complexity
K-Means Clustering	No	One word only	No	No	$O(nkt)$ k: initial clusters n: no. of documents t: iteration
Suffix Tree Clustering	Yes	Shorter but appropriate	Yes	Yes, but merging phase is not incremental	$O(n)$
Lingo	No	Longer more descriptive	Yes	No	$O(n)$
Semantic Suffix Tree	Yes	Meaningful and readable labels	Yes	Yes	$O(n)$
Improved K-Means	Yes	Based on K-means first and then on the documents linked to it	No	No	Time Consuming
Inductive Clustering	Yes	Phrases extracted from results from internal and external summary	Yes	No	Negligible with cluster titles
Fuzzy C-Mediod Clustering	No	Produce category	Yes	Yes	$O(n^2)$
Histogram-based Clustering	Yes	Matching phrases of documents	Yes	Yes	$O(n^2)$
Hierarchical Clustering	No	Most frequent terms from inside clusters	No	Yes	$O(n^2)$: single link $O(n^3)$: complete link
Semantic Hierarchical Online Clustering (SHOC)	Yes	Labels that describe clusters (extract frequent phrases and SVD technique)	Yes	Yes	$O(n)$

SVD, singular value decomposition.

STC is fast (linear to the number of documents) and incremental thus it is very useful in search results clustering process which is online postretrieval process where time is critical requirement [8]. STC clusters documents or search resultsnippets containing common phrases (sequence of words or single terms) and uses information about frequency and order of terms in the documents. STC works in two main phases namely: (i) base cluster discovery using a suffix tree and (ii) merging base clusters into proper clusters. Firstly, STC summarizes document contents and extracts phrases to be assigned then as cluster labels and thus produces concise and meaningful cluster labels [6] depending on candidate frequent phrases describing the main topic covered by the document contents. Secondly, STC assigns snippets to each of these labels to form proper clusters. Thresholds are used to manage the clustering process but tuning these thresholds is often problematic [6].

In the work described in [12], documents are also treated as strings (sequence of words) and similarity between documents is computed using string-kernel function where similarity between two documents is the number of matching subsequences. More shared substrings (not always contiguous) means more similar documents. To grouping the documents, Spectral clustering is used which is a graph-based clustering algorithm where in short, the clustering problem is a graph cut problem to isolate set of nodes from others in the collection.

In general, there are three essential steps for any WSRC method [13] listed as follows:

1. Retrieve a list $R = (r_1, r_2, \cdots, r_n)$ of n search results for the user query q.
2. Cluster R to form a list $C = (C_0, C_1, \cdots, C_m)$ of $m + 1$ clusters.
3. Label clusters.

The method described above uses each created cluster to extract a meaningful label to be assigned as a good descriptor for that cluster while in [14], for example, labels are induced first and then clustering is performed by assigning snippets to the closest preextracted label. This is the same in the Lingo "description comes first" approach which uses frequent phrases to induce distinct enough labels to cover as much topics as possible, and after that the clustering is performed by assigning each snippet to the closest label [8, 11]. Steps for "description comes first" approach are listed briefly as the following:

1. Preprocessing the input snippets by performing tokenization, stemming, and stop-words removal.
2. Extracting frequent words and phrases in the input snippets.
3. Inducing cluster labels by employing singular value decomposition (SVD).
4. Assigning snippets to each of these labels to form proper clusters.
5. Postprocessing like clusters merging and pruning.

In addition to clustering documents automatically in acceptable time, it is essentially to assign a meaningful and comprehensible label to each cluster to describe the semantic topic covered by that cluster concisely. Labelling is not a priority in traditional data

mining approaches which is mainly concerned in grouping data precisely and efficiently. While WSRC is concerned in making search results easier to brows by grouping search results in well-described clusters [6] in order to make it easier to locate the required documents and even unexpected relevant documents by reviewing certain cluster [3].

Query: **Donald Trump**		
Language	**Source**	**No. of Snippets**
English	Bing	26
	Google	40
	Wikipedia	40
	Yahoo	26
	DeuSu	40
	Base	40
	Faroo	11

No. of Sources =	7
No. of All Snippets =	223
No. of Unique Snippets =	180
No. of duplicate Snippets =	43

Figure 2 | Query used for clustering.

3. CLUSTERS LABELLING

Extracting relevant terms for labelling clusters and act as readable, meaningful and distinguishing group descriptor is a challenging process especially in WSRC where search results snippets (small portion of text) contain few terms. The high locally frequent and low globally frequent term in a cluster is typically good representative label for that cluster [3]. Terms can be weighted using local and global factors as the following:

i. **Local Factor:**

$$L_t = lg(1 + F_{Ct}) \tag{1}$$

F_{Ct} is the frequency of documents containing term t in cluster C. Logarithmic frequency is used to avoid F_{Ct} high-frequency problem.

ii. **Global Factor:**

$$G_t = lg\left(\frac{\frac{F_{Ct}}{|C|}}{\frac{F_{Rt}}{|R|}}\right) \tag{2}$$

F_{Rt} is the frequency of documents containing term t in search results R.

Label selection criterion combines local and global factors to calculate scores for terms in each cluster as the following:

$$Score_t = L_t \times G_t \tag{3}$$

For each cluster, the term with the highest score will be selected as the cluster label.

[15] used Lingo algorithm to extract frequent phrases and original terms in addition to synonym terms from WordNet lexical database, in order to induce better abstractive labels for clusters. Other external knowledge resources like Wikipedia [†]can be used to enrich the candidate label with new meaningful terms imported from the online free encyclopedia which contains a huge amount of "controlled" preclustered and manually annotated contents [4].

[16] proposed an approach to extract and combine significant *bi-grams* into *n-grams* according to term co-occurrence statistics and use the top-ranked unredundant phrases as candidate labels. To retrieve significant *bi-grams*, for each pair of words $<w, w_i>$, *strength* is computed as the following:

$$strength = \frac{freq - \bar{f}}{\sigma} \tag{4}$$

$$\bar{f} = \frac{1}{n}\sum_{1\le i\le n} freq_i$$

$$\sigma = \sqrt{\sum_{1\le i\le n}\frac{\left(freq_i - \bar{f}\right)}{n}}$$

Word pairs with *strength* value < the threshold β_0 will be discarded.

Also *spread* is computed as the following:

$$spread = \frac{\sum_{-d\le j\le d,\ j\neq 0}(p_i^j - \overline{p_i})^2}{2d} \tag{5}$$

$$\overline{p_i} = \frac{\sum_{-d\le j\le d,\ j\neq 0} p_i^j}{2d}$$

spread describes the shape of the p_i^j histogram. Small *spread* value indicates flat histogram which means that w_i can be used equivalently in almost any position around w, while large *spread* value indicates histogram with peaks which means that w_i can only be used in one (or several) specific position around w. Word pairs with *spread* value < the given threshold ρ_0 will be discarded too. The remaining word pairs are the significant *bi-grams*.

Now, *bi-grams* will be used to discover *n-grams*. Each *bi-gram* $<w, w_i>$ will be represented in a graph as a directed edge and the two words will be the vertices. A *tri-gram* "abc" is identified if the edges $a \to b, b \to c, a \to c$ exist.

n-gram is defined as $n - gram(w_1 \cdots w_n) =$

$$\left\{\begin{array}{l} edge(w_1 w_2) \ \ if\, n = 2 \\ n - gram(w_1 \cdots w_{n-1}) \ \wedge \\ \bigwedge_{i=1}^{n-1} edge(w_i w_n) \ \ if\, n > 2 \end{array}\right\}$$

where, $edge(w_1 w_2) = \left\{\begin{array}{ll} true & if(w_1, w_2) \\ false & otherwise \end{array}\right\}$

[†]http://www.wikipedia.org/

RS-ID	Title	Source	Snippet	URL
0	Donald Trump - Wikipedia	Bing	Donald John Trump (born June 14, 1946) is an American businessman, television personality, politician, and the 45th President of the United States. Trump was …	https://en.wikipedia.org/wiki/Donald_Trump
0	Donald Trump - Wikipedia	Google	Donald John Trump (born June 14, 1946) is an American businessman, television personality, politician, and the 45th President of the United States. Trump was …	https://en.wikipedia.org/wiki/Donald_Trump
0	Donald Trump - Wikipedia	Wikipedia	Donald John Trump (born June 14, 1946) is an American businessman, television personality, politician, and the 45th President of the United States. Trump was …	https://en.wikipedia.org/wiki/Donald_Trump
0	Donald Trump - Wikipedia	Yahoo	Donald John Trump (born June 14, 1946) is an American businessman, television personality, politician, and the 45th President of the United States. Trump was …	https://en.wikipedia.org/wiki/Donald_Trump
1	Make America Great Again! | Donald J Trump for President	Bing	Donald J. Trump is the very definition of the American success story, continually setting the standards of excellence in business, real estate and entertainment.	https://www.donaldjtrump.com/
1	Make America Great Again! | Donald J Trump for President	Google	Donald J. Trump is the very definition of the American success story, continually setting the standards of excellence in business, real estate and entertainment.	https://www.donaldjtrump.com/
1	Make America Great Again! | Donald J Trump for President	Yahoo	Donald J. Trump is the very definition of the American success story, continually setting the standards of excellence in business, real estate and entertainment.	https://www.donaldjtrump.com/
2	Donald Trump - CNBC - CNBC.com	Bing	Donald Trump. Republican presidential candidate Donald Trump speaks during a campaign rally at the American Airlines Center on. The nation's 45th president.	http://www.cnbc.com/donald-trump/
2	Donald Trump - CNBC - CNBC.com	Google	Donald Trump. Republican presidential candidate Donald Trump speaks during a campaign rally at the American Airlines Center on. The nation's 45th president.	http://www.cnbc.com/donald-trump/
2	Donald Trump - CNBC - CNBC.com	Yahoo	Donald Trump. Republican presidential candidate Donald Trump speaks during a campaign rally at the American Airlines Center on. The nation's 45th president.	http://www.cnbc.com/donald-trump/
3	Donald Trump | US news | The Guardian	Bing	Sean Spicer said an inquiry into Trump's baseless accusation against Obama was only in 'first chapter' and there was 'interesting news' yet to come …	https://www.theguardian.com/us-news/donaldtrump
3	Donald Trump | US news | The Guardian	Google	Sean Spicer said an inquiry into Trump's baseless accusation against Obama was only in 'first chapter' and there was 'interesting news' yet to come …	https://www.theguardian.com/us-news/donaldtrump
3	Donald Trump | US news | The Guardian	Yahoo	Sean Spicer said an inquiry into Trump's baseless accusation against Obama was only in 'first chapter' and there was 'interesting news' yet to come …	https://www.theguardian.com/us-news/donaldtrump
4	President Donald J. Trump - Biography.com	Bing	Synopsis. U.S. President, real estate mogul and former reality TV star Donald John Trump was born in 1946, in Queens, New York. In 1971, he became involved …	http://www.biography.com/people/donald-trump-9511238
4	President Donald J. Trump - Biography.com	Google	Synopsis. U.S. President, real estate mogul and former reality TV star Donald John Trump was born in 1946, in Queens, New York. In 1971, he became involved …	http://www.biography.com/people/donald-trump-9511238
4	President Donald J. Trump - Biography.com	Yahoo	Synopsis. U.S. President, real estate mogul and former reality TV star Donald John Trump was born in 1946, in Queens, New York. In 1971, he became involved …	http://www.biography.com/people/donald-trump-9511238
5	Donald Trump News and Photos | Perez Hilton	Bing	Donald Trump's ratings are in! And not unlike that of Celebrity Apprentice, the president's approval ratings can be described as "bad (pathetic)," with a dismal 39 …	http://perezhilton.com/category/donald-trump/
5	Donald Trump News and Photos | Perez Hilton	DeuSu	Donald Trump's ratings are in! And not unlike that of Celebrity Apprentice, the president's approval ratings can be described as "bad (pathetic)," with a dismal 39 …	http://perezhilton.com/category/donald-trump/
5	Donald Trump News and Photos | Perez Hilton	Google	Donald Trump's ratings are in! And not unlike that of Celebrity Apprentice, the president's approval ratings can be described as "bad (pathetic)," with a dismal 39 …	http://perezhilton.com/category/donald-trump/
5	Donald Trump News and Photos | Perez Hilton	Yahoo	Donald Trump's ratings are in! And not unlike that of Celebrity Apprentice, the president's approval ratings can be described as "bad (pathetic)," with a dismal 39 …	http://perezhilton.com/category/donald-trump/
6	Donald J. Trump (@realDonaldTrump) | Twitter	Bing	34.6K tweets • 1,993 photos/videos • 26.9M followers. Check out the latest Tweets from Donald J. Trump (@realDonaldTrump) …	https://twitter.com/realdonaldtrump
6	Donald J. Trump (@realDonaldTrump) | Twitter	Google	34.6K tweets • 1,993 photos/videos • 26.9M followers. Check out the latest Tweets from Donald J. Trump (@realDonaldTrump) …	https://twitter.com/realdonaldtrump
6	Donald J. Trump (@realDonaldTrump) | Twitter	Yahoo	34.6K tweets • 1,993 photos/videos • 26.9M followers. Check out the latest Tweets from Donald J. Trump (@realDonaldTrump) …	https://twitter.com/realdonaldtrump

Figure 3 | A list of the first 6 web search results retrieved from different web search engines (meta search) for the query: *Donald Trump* and containing a title, a URL and a snippet for each search result.

Clustering and Labelling results using (STC) Algorithm				
Cluster ID	**Size**	**No. of Phrases**	**Cluster Label**	**Snippets IDs**
0	14	3	Purdue University Indianapolis IUPUI, Indiana University, Donald J. Trump Foundation	60,67,92,94,98,102,104,106,111,115,119,126,158,178
1	40	2	Latest, Donald Trump News	3,5,6,8,12,13,18,36,39,41,44,46,50,63,65,66,72,74,76, 78,81,83,85,95,97,105,108,110,112,116,121,125,128, 131,149,151,159,167,169,178
2	25	1	United States	0,10,15,19,20,21,30,45,49,58,62,67,75,79,80,83,86,116, 117,124,141,147,150,156,166
3	57	1	President	0,1,2,4,9,10,12,13,14,15,19,21,28,30,35,38,42,44,45,46, 47,49,51,55,56,58,62,64,71,73,75,77,79,80,81,91,99,103, 108,113,116,122,123,124,129,134,136,137,141,147,148, 150,154,157,158,166,169
4	10	3	Born, Donald John Trump, 1946	0,4,11,19,21,29,30,92,107,170
5	26	1	Donald Trump's	5,10,12,17,22,25,34,38,43,44,46,58,65,86,113,133,134, 140,157,162, 165,169,172,173,174,179
6	11	2	Campaign, Donald Trump Presidential Campaign	2,14,25,53,96,100,113,130,154,158,165
7	10	2	Candidate, Presidential Candidate Donald Trump	2,47,81,88,108,127,139,153,158,165
8	16	1	45th President	0,2,10,15,19,21,30,42,49,58,62,71,75,79,108,141
9	23	2	Videos, Photos	5,6,8,18,36,46,50,63,66,70,72,74,83,87,93,95,118,121, 125,127,128,163,169
10	17	1	Donald J. Trump	1,4,6,7,11,15,57,60,70,94,98,102,106,111,115,117,119
11	5	3	People, Protest, Took Part in Marches	37,86,172,176,179
12	10	1	New York	4,11,14,28,29,51,57,85,153,171
13	11	1	Television Personality	0,49,53,58,81,107,108,127,140,141,160
14	11	1	American Businessman	0,14,21,40,49,53,58,79,82,141,150
15	37	*	Other Topics	16,23,24,26,27,31,32,33,48,52,54,59,61,68,69,84,89,90, 101,109,114,120,132,135,138,142,143,144,145,146,152, 155,161,164,168,175,177
Total	16	323	25	16

No. of All Snippets in Clusters =	323
No. of Unique Snippets in Clusters =	180
No. of duplicate Snippets in Clusters =	143
Avg. no. of Snippets in a Cluster =	20.19

Avg. no. of Words per Phrase =	2.12
Avg. no. of Phrases per Label =	1.67

Figure 4 | Clustering and labelling results using the classical Suffix Tree Clustering (STC) algorithm for the query: *Donald Trump*.

Depth-first traversal to all nodes results in extracting all the *n-grams*. After that, redundant *n-grams* need to be eliminated and only unredundant *n-grams* will be used as candidate cluster labels. Removing the redundancy in *n-grams* can be performed by applying *remove-or-merge* process.

Let $ts(p)$ be the term set of *n-gram* p, $ss(p)$ be the sentence set of p, and ω_0 be a threshold. The *remove-or-merge* condition is defined as:

$$\left\{ \begin{array}{l} if \left| \dfrac{ts(p_i) \cap ts(p_j)}{min\{ts(p_i), ts(p_j)\}} \right| = 1 \ \& \left| \dfrac{ss(p_i) \cap ss(p_j)}{ss(p_i) \cup ss(p_j)} \right| \geq \omega_0, \ delete\ shorter\ gram \\ if \left| \dfrac{ts(p_i) \cap ts(p_j)}{ts(p_i) \cup ts(p_j)} \right| \geq \omega_0 \ \& \left| \dfrac{ss(p_i) \cap ss(p_j)}{ss(p_i) \cup ss(p_j)} \right| \geq \omega_0, \ merge\ p_i\ and\ p_j \end{array} \right\}$$

Each candidate cluster label p is ranked by its *significance* ($Sig(p)$) as the following:

$$Sig(p) = tfidf(p) \times boost(p) \quad (6)$$

$$tfidf(p) = tf(p) \times log\left(1 + \frac{N}{df(p)}\right)$$

$$boost(p) = \left\{ \begin{array}{l} c^{|p|} - base, |p| \leq 8 \\ 5, |p| > 8 \end{array} \right\}$$

$boost(p)$ is a boost factor of phrase p, c and $base$ are constants ($c = 1.25$ and $base = 0.5$).

The top M candidate cluster labels are selected to construct base clusters. All snippets containing the same label (phrase) are aggregated in a base cluster labelled by the phrase.

Even though it is generated automatically, evaluation of clusters labels may be better to be conducted against manually created gold standard data where human annotators are asked to identify the fittest cluster given a cluster label [17].

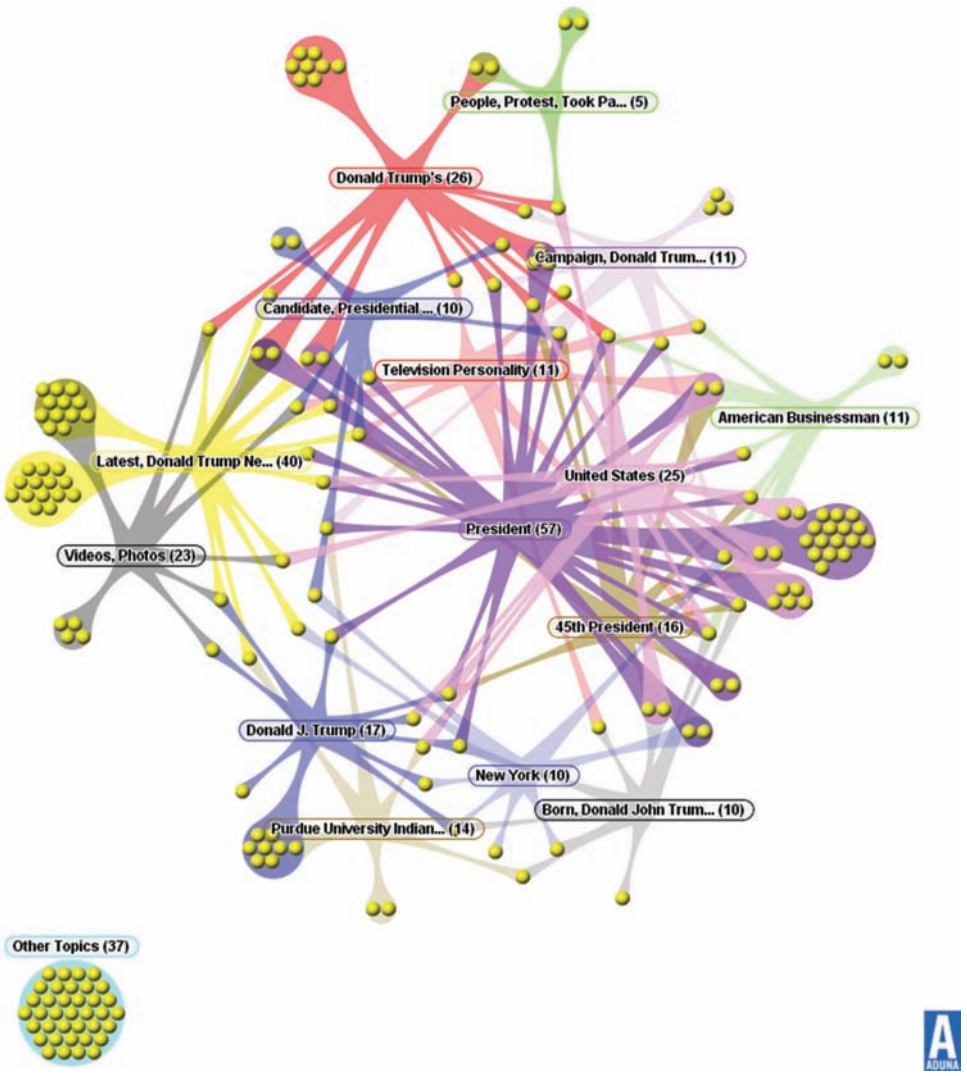

Figure 5 | Clusters visualisation for the resulted clusters produced from the Suffix Tree Clustering (STC) algorithm for the query: *Donald Trump*.

4. STC CLUSTERS LABELLING ENHANCEMENT

The enhancement process is founded on the idea of using the existing labels nominated by the standard STC algorithm. The original labels and/or clusters will be modified and combined so that it become more concise and descriptive. To this end, the propose methodology will be conducted on the original STC algorithm to produce an enhanced version of the classical STC algorithm. The proposed methodology employed a deeper linguistic analysis and more robust techniques (as seen in Algorithm 1) than that used in other research works like, for example, the work described in [18].

Once the raw original labels are induced using STC algorithm all the cluster label phrases and clusters will be processed with respect to the Algorithm 1. The aim of this algorithm is to enhance f1, f2, f3, f4, and f5 (see Section 5 for details) by refining and reformulating both of labels and clusters. The major steps are described Algorithm 1.

5. CLUSTERS LABELLING QUALITY EVALUATION

Labels for web search results clusters should be discriminatory and carefully describe the contents of individual clusters. Low-quality labelling for web search results clusters may confuse the user and mislead him during navigation through clusters and thus negatively affect the whole process aiming to meet the user information needs [18]. In this section, a discussion concerning the evaluation of the generated labels for the produced clusters of web search results is presented. This is important for generating descriptive and precise labels for clusters and/or conducting a comparison concerning the descriptiveness of different labelling techniques.

Clusters labelling quality measures can be conducted as an external measure according to the source of the "validity criteria." External measure compares the clustering algorithm's results against external, manually, or automatically, prelabelled results in order to compare the difference between the two results. Many labelling

Algorithm 1 Labels Enhancement and Clusters Refinement

INPUT: Set of Labels $L = \{l_1, l_2, \ldots, l_n\}$
INPUT: Set of Clusters $C = \{c_1, c_2, \ldots, c_n\}$
OUTPUT: Set of Enhanced Labels $L' = \{l'_1, l'_2, \ldots, l'_n\}$
OUTPUT: Set of Refined Clusters $C' = \{c'_1, c'_2, \ldots, c'_m\}$

procedure PHRASES PREPARATION AND PREPROCESSING
    Phrases preparation and preprocessing:
    Lemmatise Phrases, Part Of Speech Tagging, Find Synonyms, Resolve Proper-nouns Ambiguity
    for all $l_k \in L$ *Associated with* $c_k \in C$ **do**
        for all $p_i \in l_k$ **do**
            for all $t_{ij} \in Tokenise_{word}(p_i), j = 1...length(p_i)$ **do**
                Find the dictionary reduced form i.e Lemma(t_{ij}) *using Snowball Lemmatiser*
                Replace t_{ij} *by Lemma*(t_{ij}) *for all of its occurrences in all phrases*
                Find POST(t_{ij}) *using TAIParse Tagger*
                Find Synonym(t_{ij}) *from WordNet*
                Create an Eqiuvilent class and
                Append t_{ij} *and all of its synonyms. Name it as* $Equiv_{ij}$
                Replace all terms $\in Equiv_{ij}$ *by the commonest word* $\in Equiv_{ij}$ *or*
                commonest (having highest DF) word (term) $t_{ij} \in p_i$. *This will increase f*3
                Normalise(t_{ij}) *by standardising all word variations into equivalent classes*
            for all t_i *which is Proper Nouns* **do**
                Find DisAmb(t_{ij}) *from SOUNDEX to resolve possible phonetic variations*

procedure PHRASES REFINE
    if t_i *is Stop Word (in LIST of Stop Words)* **then**
        if t_i *is not at the beginning or ending of the* p_i **then**
            Remove t_{ij} *from* p_i

    Merge clusters if have common label
    remove phrase from one cluster and keep it in the second one
    BUT move associated documents consequently (documents must follow its phrase)
    AND neglect replicated documents within a cluster

quality measures have been proposed in different contexts in the literature. [18] introduced a new metric to evaluate the quality of clusters labels using a comparative evaluation strategy. The authors in [18] argued that, to be responsible, clusters labelling evaluation should take into consideration the following five parameters:

1. *Comprehensibility (f1)*: A cluster label should give a clear interpretation for the contents of a cluster to the user. It can be formally defined as $\forall c \in C\, \forall p \in lc : P \in L(G)^p > 1$, where lc is the cluster label of cluster c, and $L(G)$ define a formal language identifying noun phrases (a word or group of words containing a noun and functioning in a sentence as subject, object, or prepositional object.).

$$f1\,(p) = NP\,(p) * Penalty\,(p) \tag{7}$$

$$NP\,(p) = \begin{Bmatrix} 1, & if\ P \in L(G) \\ 0, & Otherwise \end{Bmatrix}$$

$$Penalty\,(p) = \begin{Bmatrix} exp\dfrac{-(|p| - |p|_{opt})^2}{2 \cdot d^2}, & if\ |p| > 1 \\ 0.5, \quad Otherwise \end{Bmatrix}$$

The exponential expression in Equation (7) is used to penalize too short or too long phrases by setting $|p|_{opt} = 4$ and $d = 8$. [19].

2. *Descriptiveness (f2)*: All documents in a cluster should contain the label associated with that cluster. It can be formally defined as: $\forall c \in C\, \exists p \in l_c\, \forall p' \in_{p' \notin l_c} P_c : \ll df_c(p)$. Where P_c is the set of phrases in cluster c. p' is the complement value of p, $df_c(p)$ represents the number of documents in a cluster containing phrase p.

$$f2\left(c,p\right)=1-\frac{1}{|P_c\setminus l_c|}\sum_{\substack{p'\in P_c\\ p'\notin l_c}}\frac{df_c\left(p'\right)}{df_c\left(p\right)} \tag{8}$$

3. *Discriminative Power (f3)*: A cluster label should only exist, exclusively, in documents from its associated cluster. It can be formally defined as:

$$\forall c_i, c_{jc_i\neq c_j}\in C\,\exists P\in l_c\,:\,\frac{df_{ci}\left(p\right)}{|c_i|}\ll\frac{df_{cj}\left(p\right)}{|c_j|}$$

$$f3\left(c_j,p\right)=1-\frac{1}{k-1}\sum_{\substack{c_i\in C\\ c_i\neq c_j}}\frac{|c_j|\cdot df_{ci}\left(p\right)}{|c_i|\cdot df_{cj}\left(p\right)} \tag{9}$$

Where c_i and c_j are two clusters while $df_c(p)$ represents the number of documents in a cluster containing the phrase p.

4. *Uniqueness (f4)*: Each cluster label should be uniquely associated with one cluster. It can be formally defined as: $\forall c_i, c_{jc_i\neq c_j}\in C\,:\,l_{c_i}\cap l_{c_j}=\varnothing$

$$f4\left(c_j,p\right)=1-\frac{1}{k-1}\sum_{\substack{c_i\in C\\ c_i\neq c_j}}\frac{|p\cap l_{c_i}|}{|p\cup l_{c_j}|} \tag{10}$$

Where p is a phrase and l_c is the label associated with a cluster.

5. *Nonredundancy (f5)*: Cluster labels can not be synonymous (having the same or nearly the same meaning). It can be formally defined as$\forall c\in C\forall p,p'\in l_c\,:\,p$ and p' are not synonymous $[p\neq p']$.

$$f5\left(c,p\right)=1-\frac{1}{|l_c|-1}\sum_{\substack{p'\in l_c\\ p'\neq p}}Syn\left(p,p'\right) \tag{11}$$

Where $Syn\,:\,p\times p\rightarrow\{0,1\}$.

Label relevancy: Relevance of a phrase with respect to a cluster: All constraints can be combined into a single criterion:

$$rel\left(c,p\right)=\sum_{i=1}^{|\mathcal{F}|}w_i\cdot f_i\left(c,p\right) \tag{12}$$

Where *wi* is a weighting factor and $\mathcal{F}=\{f|1...5\}$, namely:

- $f1$: Comprehensibility
- $f2$: Descriptiveness
- $f3$: Discriminative Power
- $f4$: Uniqueness
- $f5$: Nonredundancy

Clusters labelling quality measures can be categorized as (i) external, (ii) internal, and (iii) relative measures according to the source of the "validity criteria." External measure compares the clustering algorithm's results against external, manuallyor automatically, preclustered results in order to disclose the difference between the two results. While internal measure employs functions to assess the similarity between cluster's documents in addition to the dissimilarity between resulted clusters without referring to any external information. Relative measure assesses the results by comparing them against results from different algorithms, or compares the results of the same algorithm but under different conditions like different thresholds [20].

6. RESULTS AND DISCUSSION

One of the challenges of work on WSRC is the lack of "ground truth" data. In some cases it is possible to construct such data by hand however this still entails subjectivity and requires considerable resources (to the extent that it is not possible to construct significant benchmark data).

To act as a focus for the work described in this paper, the top-ranked web search results automatically *clustered* into relatively small thematic collections of documents for the query *Donald Trump* (see Figure 4). These clustered web search results retrieved from Carrot[2] which is an open source WSRC engine (available on: http://search.carrot2.org/stable/search) using STC algorithm.

To evaluate the proposed methodology the authors compared the reformulated clusters and their enhanced labels with the original clusters and labels generated from the classical STC algorithm. For evaluation, the authors used clusters labelling performancemeasure considered five parameters (as discussed in Section 5). We also had to forget, for the purpose of the evaluation, the numerical "intensity" of the computed values for f1: Comprehensibility, f2: Descriptiveness, f3: Discriminative Power, f4: Uniqueness, and f5: Nonredundancy.

The results are presented in tabular form and show the performance of the proposed methodology to enhance the STC algorithm with respect to the classical STC algorithm. The evaluation shows that the proposed methodology to enhance the STC algorithm performs well with respect to the quality of the enhanced labels for clusters. Inspection of the recorded results in Figure 7 indicates that (i) The proposed methodology achieved better performance and the overall average recorded values for the used performance measure (f6) was 0.921. (ii) Number of clusters was decreased from 15 to clusters only. (iii) Number of duplicated results was decreased from 143 to 121 only, and (iv) average number of phrases per label was increased from 1.67 to 2.00 phrases.

7. CONCLUSIONS

In this paper the authors described the proposed methodology for enhancing the classical STC algorithm for clustering web search results. The operation of the proposed methodology was illustrated and evaluated. The objective of this research was to deploydeep linguist analysis techniques for the enhancement of phrase labels that will in turn allow for the reformulation of the structure of web search results clusters and thus produce better performance for web search engines and achieve a better end-user satisfaction. The proposed methodology used the existing labels nominated by

Clustering and Labelling results using (Enhanced STC) Algorithm				
Cluster ID	**Size**	**No. of Phrases**	**Cluster Label**	**Snippets IDs**
0	14	3	Purdue University Indianapolis IUPUI, Indiana University, Donald John Trump Foundation	60,67,92,94,98,102,104,106,111,115,119,126,158,178
1	40	2	Latest Donald Trump News	3,5,6,8,12,13,18,36,39,41,44,46,50,63,65,66,72,74,76, 78,81,83,85,95,97,105,108,110,112,116,121,125,128, 131,149,151,159,167,169,178
2	35	2	United States of America, New York City	0,4,10,11,14,15,19,20,21,28,29,30,45,49,51,57,58,62, 67,75,79,80,83,85,86,116,117,124,141,147,150,153, 156,166, 171
3	57	1	45th President	0,1,2,4,9,10,12,13,14,15,19,21,28,30,35,38,42,44,45, 46,47,49,51,55,56,58,62,64,71,73,75,77,79,80,81,91, 99,103,108,113,116,122,123,124,129,134,136,137, 141,147,148,150,154,157,158,166,169
4	59	2	Donald John Trump, American Businessman	0,1,4,5,6,7,10,11,12,14,15,17,19,21,22,25,29,30,34,38, 40,43,44,46,49,53,57,58,60,65,70,79,82,86,92,94,98, 102,106,107,111,113,115,117,119,133,134,140,141, 150,157,162, 165,169,170,172,173,174,179
5	11	2	Donald John Trump Presidential Campaign	2,14,25,53,96,100,113,130,154,158,165
6	10	1	Presidential Candidate Donald John Trump	2,47,81,88,108,127,139,153,158,165
7	33	2	Television Personality, Videos and Photos	0,5,6,8,18,36,46,49,50,53,58,63,66,70,72,74,81,83,87, 93,95,107,108,118,121,125,127,128,140,141,160,163, 169
8	5	3	People, Protest, Took Part in Marches	37,86,172,176,179
9	37	*	Other Topics	16,23,24,26,27,31,32,33,48,52,54,59,61,68,69,84,89, 90,101,109,114,120,132,135,138,142,143,144,145, 146,152,155,161,164,168,175,177
Total	10	301	18	16

No. of All Snippets in Clusters =	301
No. of Unique Snippets in Clusters =	180
No. of duplicate Snippets in Clusters =	121
Avg. no. of Snippets in a Cluster =	30.10
Avg. no. of Words per Phrase =	2.78
Avg. no. of Phrases per Label =	2.00

Figure 6 | Clustering and Labelling results using the enhanced Suffix Tree Clustering (STC) algorithm for the query: Donald Trump.

the original STC algorithm and adapted that labels and/or clusters to be more concise and descriptive. The propose methodology was conducted on the original STC algorithm to produce an enhanced version of the classical Suffix Tree Clustering algorithm. The enhanced algorithm was experimented and the produced clusters and labels were compared and evaluated with respect to the classical STC algorithm using clusters labelling performance measure considered five parameters f1: Comprehensibility, f2: Descriptiveness, f3: Discriminative Power, f4: Uniqueness, and f5: Nonredundancy. The recorded results indicated that the new enhanced labels outperformed the original labels and the overall performance has been enhanced. The results shown that better performance was achieved (f6 = 0.921), clusters were decreased (from 15 to 9 clusters only), duplicated web search results were decreased (from 143 to 121 only), and average number of phrases per label was increased (from 1.67 to 2.00 phrases).

The promising results obtained so far indicate that (i) it is possible to capture the clusters structure (ii) it is possible to enhance the produced labels from STC algorithm to improve the overall performance by producing more comprehensive and descriptive labels for clusters and thus the user will be able to preview and navigate easily and fast.

Future work will initially be directed at the adoption of deeper linguistic approaches and data mining techniques to enhance other WSRC algorithms like Lingo and K-mean. The intention is also to increase the size of our dataset.

Clustering and Labelling results using (STC) Algorithm								
Cluster ID	**No. of Phrases**	**Cluster Label**	f1	f2	f3	f4	f5	f6
0	3	Purdue University Indianapolis IUPUI, Indiana University, Donald J. Trump Foundation	0.990	0.786	0.851	0.875	1.000	0.900
1	2	Latest, Donald Trump News	0.496	0.950	0.901	0.917	1.000	0.853
2	1	United States	0.969	0.960	0.950	0.958	1.000	0.968
3	1	President	0.500	0.982	0.950	0.958	0.000	0.678
4	3	Born, Donald John Trump, 1946	0.331	0.700	0.851	0.875	0.667	0.685
5	1	Donald Trump's	0.969	0.962	0.950	0.958	0.000	0.768
6	2	Campaign, Donald Trump Presidential Campaign	0.750	0.818	0.901	0.917	1.000	0.877
7	2	Candidate, Presidential Candidate Donald Trump	0.750	0.800	0.901	0.917	1.000	0.873
8	1	45th President	0.500	0.938	0.950	0.958	0.000	0.669
9	2	Videos, Photos	0.500	0.913	0.901	0.917	1.000	0.846
10	1	Donald J. Trump	0.992	0.941	0.950	0.958	0.000	0.768
11	3	People, Protest, Took Part in Marches	0.167	0.400	0.851	0.875	1.000	0.659
12	1	New York	0.969	0.900	0.950	0.958	1.000	0.956
13	1	Television Personality	0.969	0.909	0.950	0.958	1.000	0.957
14	1	American Businessman	0.969	0.909	0.950	0.958	1.000	0.957
		Average	**0.721**	**0.858**	**0.917**	**0.931**	**0.711**	**0.828**
		Max	0.992	0.982	0.950	0.958	1.000	0.968
		Min	0.167	0.400	0.851	0.875	0.000	0.659

Clustering and Labelling results using (Enhanced STC) Algorithm								
Cluster ID	**No. of Phrases**	**Cluster Label**	f1	f2	f3	f4	f5	f6
0	3	Purdue University Indianapolis IUPUI, Indiana University, Donald John Trump Foundation	0.999	0.786	0.901	0.958	1.000	0.929
1	2	Latest Donald Trump News	0.999	0.950	0.901	0.917	1.000	0.953
2	2	United States of America, New York City	0.999	0.980	0.950	0.958	1.000	0.978
3	1	45th President	0.999	0.986	0.950	0.958	1.000	0.979
4	2	Donald John Trump, American Businessman	0.999	0.984	0.950	0.958	1.000	0.978
5	2	Donald John Trump Presidential Campaign	0.999	0.818	0.950	0.917	1.000	0.937
6	1	Presidential Candidate Donald John Trump	0.999	0.984	0.950	0.917	1.000	0.970
7	2	Television Personality, Videos and Photos	0.750	0.965	0.901	0.917	1.000	0.906
8	3	People, Protest, Took Part in Marches	0.167	0.400	0.851	0.875	1.000	0.659
		Average	**0.879**	**0.873**	**0.923**	**0.931**	**1.000**	**0.921**
		Max	0.999	0.986	0.950	0.958	1.000	0.979
		Min	0.167	0.400	0.851	0.875	1.000	0.659

Figure 7 | Classical Vs enhanced Suffix Tree Clustering (STC).

REFERENCES

[1] R.K. Roul, S.K. Sahay, Cluster labelling using chi-square-based keyword ranking and mutual information score: a hybrid approach, Int. J. Intell. Syst. Des. Comput. 1(1–2) (2017), 145–167.

[2] H. Chim X. Deng. Efficient phrase-based document similarity for clustering. IEEE Trans. Knowl. Data Eng. 20(9) (2008), 1217–1229.

[3] H.-M. Li, C.-X. Sun, K.-J. Wang, Clustering web search results using conceptual grouping, in 2009 International Conference on Machine Learning and Cybernetics, vol. 3, IEEE, Red Hook, NY, 2009, pp. 1499–1503.

[4] D. Carmel, H. Roitman, N. Zwerdling, Enhancing cluster labeling using Wikipedia, in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, 2009, pp. 139–146.

[5] G. Kr Yadav, A. Kumar, A described feasibility analysis on web document clustering, Int. J. Res. Stud. Sci. Eng. Technol. 2 (2015), 1–13.

[6] U. Bharambe, A. Kale, Landscape of web search results clustering algorithms, in: S. Unnikrishnan, S. Surve, D. Bhoir (Eds.), Advances in Computing, Communication and Control, Springer, Switzerland, 2011, pp. 95–107.

[7] H. Agrawal, S. Yadav, Search engine results improvement–a review, in IEEE International Conference on Computational Intelligence Communication Technology, Piscataway, NJ, February 2015, pp. 180–185.

[8] S. Kopidaki, P. Papadakos, Y. Tzitzikas, STC+ and NM-STC: two novel online results clustering methods for web searching, in Web Information Systems Engineering (WISE 2009), Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 523–537.

[9] T. Rani, A. Goyal, Survey of clustering techniques for information retrieval in data mining, Int. J. Sci. Eng. Technol. Res. 4(4) (2015), 738–740.
[10] M. Waseem Khan, H.M. Shahzad Asif, Y. Saleem, Semantic based cluster content discovery in description first clustering algorithm. Mehran University Research J. Eng. Technol. 36(1) (2017), 1–6.
[11] S. Osinski, D. Weiss. A concept-driven algorithm for clustering search results, IEEE Intell. Syst. 20(3) (2005), 48–54.
[12] Q. Shi, X. Qiao, X. Guangquan, Using string kernel for document clustering, Int. J. Inf. Technol. Comp. Sci. 2 (2010), 40–46.
[13] A. Di, R. Navigli, Clustering web search results with maximum spanning trees, in Congress of the Italian Association for Artificial Intelligence, Springer, 2011, pp. 201–212.
[14] H.D. Abdulla, V. Snasel, Using singular value decomposition (svd) as a solution for search result clustering, in International Conference on Innovations in Information Technology, IEEE, Piscataway, NJ, 2008, pp. 302–306.
[15] A. Sameh, A. Kadray, Semantic web search results clustering using lingo and wordnet, Int. J. Res. Rev. Comput. Sci. 1(2) (2010), 71–76. http://search.proquest.com/openview/d2be52a63fc5a2f4a20f0603250e8d56/1?pq-origsite=gscholar&cbl=276284.
[16] Y. Zhang, B. Feng, A co-occurrence based hierarchical method for clustering web search results, in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, IEEE, Piscataway, NJ, 2008, pp. 407–410.
[17] A. Aker, E. Kurtic, A.R. Balamurali, M. Paramita, E. Barker, M. Hepple, R. Gaizauskas, A Graph-Based Approach to Topic Clustering for Online Comments to News, Springer International Publishing, Cham, Switzerland, 2016, pp. 15–29.
[18] R. Mahalakshmi, V.L. Praba, enhancing the labelling technique of suffix tree clustering algorithm, Int. J. Data Min. Know. Manag. Process. 4 (2014), 41–50.
[19] D. Weiss, Descriptive clustering as a method for exploring text collections, PhD thesis, Citeseer, 2006.
[20] T. Velmurugan, T. Santhanam, Clustering mixed data points using fuzzy c-means clustering algorithm for performance analysis, Int. J. Comput. Sci. Eng. 2 (2010), 3100–3105.