Automatic Metadata Generation By Clustering Extracted Representative Keywords From Heterogeneous Sources

Ali Ridho Barakbah

Abstract—In the information retrieval, the generation of important words for metadata space creation is very important to extract representative information from sources. The extraction of important words from a document source which are derived from the intensity of term currently might not represent the original source. In this paper, we propose a new approach to automatically generate a representative metadata by applying a clustering in order to extract representative keywords from heterogeneous sources. The proposed approach consists of three stages: (1) Aggregate Keyword Extraction, (2) Automatic Source Filter, and (3) Representative Keyword Generation. First of all, we extract an aggregate metadata from the all sources of the documents. Secondly, we provide an automatic mechanism to get the selected aggregate metadata by filtering out the sources and acquiring the representative sources by using a set of classifying words. Thirdly, we promote the selected aggregate metadata to be representative metadata. We apply our Hierarchical K-Means to cluster the extracted keywords in order to generate the representative keywords to realize the representative metadata. To perform the applicability of our proposed approach for automatic metadata generation, we conduct an experiment with information sources of Sidoarjo mud flow consisting of 60 English article sources related to Sidoarjo mud flow. The experimental result performs effectiveness of the proposed approach to generate the representative metadata and reduce drastically the metadata space of the keywords.

Index Term—keyword extraction, representative keyword, metadata generation, Sidoarjo mud flow.

1 Introduction

The World Wide Web has become a significant source of information. As the cost of storage devices continues to decrease there is tremendous growth in databases of all sorts (relational, graphical, and textual). Knowledge intensive organizations have vast array of information contained in large document repositories [3]. This explosive growth has led to huge, fragmented, and unstructured document collections.

Although it has become easier to collect and store information in document collections, it has become increasingly difficult to retrieve relevant information from these large document collections. Such a large amount of information serves as a huge information repository for organizations. How to help users find their required information is the central task of any information retrieval (IR) system or search engine [6]. Various techniques have been used by researchers to address the issue of improving retrieval performance.

Retrieval performance of an IR system can be affected by many factors: the ambiguity of query terms, unfamiliarity with system features, as well as ranking/matching function [5]. Another very important factor that is often overlooked by most researchers is the keywords relating to document representation. These important words are considered to make such as document classification, ranked articles, as well as creation of metadata space. The meta-level knowledge based on important words from documents can be realized as a creative environment in new research fields by integrating information resources in various re-

• *Ali Ridho Barakbah is with Electronic Engineering Polytechnic Institute of Surabaya, Knowledge Engineering Research Group. (email:nanang@eepis-its.edu, Jl. Raya ITS Keputih Sukolilo, telp:031-5947280/ext:4109, fax:031-5946114)*

Manuscript was received 3 December 2011.

Ali Ridho Barakbah, "Automatic Metadata Generation by Clustering Extracted Representative Keywords from Heterogeneous Sources",
Industrial Electronics Seminar (IES) 2011, October 26, 2011, Surabaya, Indonesia, and selected for journal publication in the Journal of Emitter, Vol. 2, No. 2, 2012.

search fields such as cultural, social and natural sciences [7]. It is this generation of important words from documents for metadata space creation that we focus most of our discussion on.

2 OBJECTIVE

Some of approaches already proposed how to get the important words from a document, such as TF-IDF, weighted approach of document, Nave Bayes approach, Genetic Algorithm approach, etc. These currently approaches emphasized by scoring the intensity of word appearance in the documents. The extraction of important words using these approaches often does not satisfy user requirements. This problem is caused by a word that intensely appears in a document can be selected as an important word, even it is a representative word of that document.

The selection of important keywords in a document already was discussed by Sasaki et al. in the creation of metadata space [4]. Human intervention was involved in he metadata space creation [4] to select manually the important words from document. However, the manual selection by human perception can not be analytically perceived and supposed as representative selection. For example, in the case of Sidoarjo mudflow article in Wikipedia said Mud volcano systems are common on Earth, including on Java island and particularly in East Java province. One person may think that earth is keyword in the document, but another person could have said just the opposite. The correctness of both opinions can be scrutinized because of different person preferences.

The other problem is, moreover, only 1 document was included keyword extraction. It brings difficulty when the representative document for keyword extraction is more than one document. For example, these following documents mentioned about definition of mud.

1. Wikipedia said "*mud is a liquid or semi-liquid mixture of water and some combination of soil, silt, and clay...*"
2. Oxford Pocket Dictionary said "*mud is soft, sticky matter resulting from the mixing of earth and water*"
3. Dictionary.com said "*mud is a mixture of chemicals and other substances pumped into a drilling rig chiefly as a lubricant for the bit and shaft*"

If only 1 document involved extracting the representative keywords, it will be difficult to choose the correct one, because all definition of each documents are correct but written in the different language expressions.

Trying to deal with these problems, in this research we propose another approach how to extract the representative keywords from heterogeneous sources by identifying similarity of words among documents. Our proposed approach can extract representative keywords automatically from heterogeneous sources. This proposed approach will apply clustering mechanism to detect the closeness of keywords in the feature space.

3 BASIC CONCEPT

Here we describe fundamental concepts of our proposed approach for clustering sources and keywords. We start the description of clustering concept, and the most simple clustering algorithm using K-means. The description ends with our approach for K-means optimization.

3.1 Clustering

Clustering is an effort to classify similar objects in the same groups. Cluster analysis constructs good cluster when the members of a cluster have a high degree of similarity to each other (internal homogeneity) and are not like members of other clusters (external omogeneity). It means process to define a mapping f:$D \rightarrow C$ from some data D=$\{d1, d2, .., dn\}$ to some clusters C=$\{c1, c2, .., cn\}$ on similarity between di. There many applications of clustering diverse in many fields, such as data mining, pattern recognition, image classification, biological sciences, marketing, city-planning, document retrievals, etc.

3.2 K-means Algorithm

The most well known methods for clustering is Kmeans developed by Mac Queen in 1967. The simplicity of K-means made this algorithm

Ali Ridho Barakbah, "Automatic Metadata Generation by Clustering Extracted Representative Keywords from Heterogeneous Sources",
Industrial Electronics Seminar (IES) 2011, October 26, 2011, Surabaya, Indonesia, and selected for journal publication in the Journal of Emitter, Vol. 2, No. 2, 2012.

used in various fields. K-means is a partition clustering method that separates data into k mutually excessive groups. By iterative such partitioning, K-means minimizes the sum of distance from each data to its clusters. K-means method is very popular because of its ability to cluster huge data, and also outliers, quickly and efficiently. It remains a basic framework for developing numerical or conceptual clustering systems because various possibilities of distance and prototype choice.

K-means algorithm can be described as follows:

1. Initiate its algorithm by generating random initial cluster centers c_k.
2. Calculate the distance $d(x, c)$ between vector x_i to cluster center c_k. Euclidean distance can be used to express the distance.
3. Separate x_i into s_k which has minimum $d(x, c)$.
4. Determine the new cluster centers defined as:

$$c_i = \frac{1}{p} \sum_{(j=1)}^{p} m(s_i, j), where p = n(s_i) \quad (1)$$

5. Go back to step 2 until $C_i = C_i - 1$.

3.3 Hierarchical K-means

However, K-means algorithm is very sensitive in initial starting points. K-means generates initial cluster randomly. When random initial starting points close to the final solution, K-means has high possibility to find out the cluster center. Otherwise, it will lead to incorrect clustering results. Because of initial starting points generated randomly, K-means does not guarantee the unique clustering results. K-means method is difficult to reach global optimum, but only in local minimum.

In this proposed research, we will use Hierarchical K-means to optimize initial centroids for K-means [2]. This algorithm is better used for the complex clustering cases with large numbers of data set and many dimensional attributes. Hierarchical K-means bargains the advantage of K-means algorithm in speed and hierarchical algorithm in precision.

It utilizes all the clustering results of K-means in several times, even though some of them reach the local optima. Then, the result transformed by combining with Hierarchical algorithm in order to determine the initial centroids for K-means.

The algorithm is described as follows:

1. Set $X = \{xi|i = 1, ..., r\}$ as each data of A, where $A = \{ai|i = 1, ..., n\}$ is attribute of n-dimensional vector.
2. Set K as the predefined number of clusters.
3. Determine p as numbers of computation
4. Set $i = 1$ as initial counter
5. Apply K-means algorithm.
6. Record the centroids of clustering results as $C_i = \{c_{ij}|j = 1, ..., K\}$
7. Increment $i = i + 1$
8. Repeat from step 5 while $i < p$.
9. Assume $C = \{C_i|i = 1, ..., p\}$ as new data set, with K as predefined number of clusters
10. Apply hierarchical algorithm
11. Record the centroids of clustering result as $D = \{d_i|i = 1, ..., K\}$

Then, $D = \{d_i|i = 1, ..., K\}$ considered as initial cluster centers for K-means clustering.

4 Proposed Approach

In this paper, we propose a new approach to automatically generate a representative metadata by extracting representative keywords from heterogeneous sources. The metadata consists of a set of concept terms with their each feature term, as shown in Figure 1. It can be divided into 3 stages those are: (1) Aggregate Keyword Extraction, (2) Automatic Source Filter, and (3) Representative Keyword Generation.

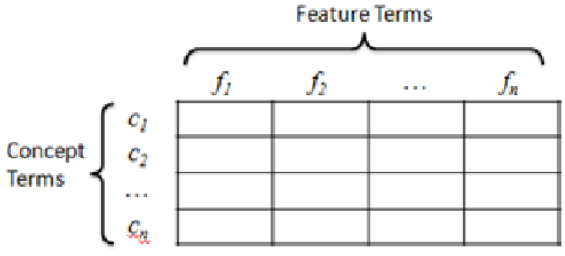

Figure 1. A generated metadata from heterogeneous sources consisting of a set of concept terms and feature terms

Ali Ridho Barakbah, "Automatic Metadata Generation by Clustering Extracted Representative Keywords from Heterogeneous Sources",
Industrial Electronics Seminar (IES) 2011, October 26, 2011, Surabaya, Indonesia, and selected for journal publication in the Journal of Emitter, Vol. 2, No. 2, 2012.

4.1 Aggregate Keyword Extraction

First of all, we extract an aggregate metadata from the all sources of the documents. In this stage, large amount of documents can be better to extract the representative keywords. We applied TF-IDF algorithm to extract the aggregate keyword from the all sources, as shown in Figure 2.

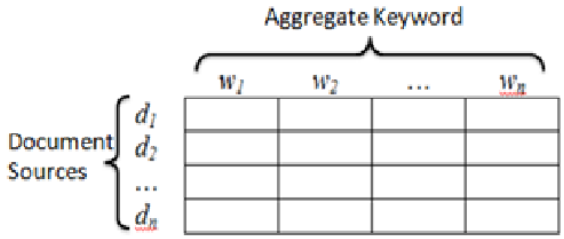

Figure 2. Aggregate metadata for the all document sources

$D = \{d1, d2, ..., dn\}$ is the collected document sources and $W = \{w1, w2, ..., wn\}$ is an aggregate keyword for each source extracted by TF-IDF algorithm.

4.2 Automatic Source Filter

Filtering out irrelevant source data set is an important issue in document mining [1]. Because the large sources generate huge aggregate metadata, it is very important to filter out the irrelevant sources to reduce a complexity of the metadata creation. We provide an automatic mechanism to filter out the sources and acquire the representative sources by using classifying words. First of all, we determine a set of classifying words to avoid the emptiness of value in high dimensional-attributes. Let $CW = \{cw1, cw2, ..., cwn\}$ be classifying words. Then, we apply an inner-product for CW to W in order to get the representative sources, but we do not make a rank of them. Then, we apply our Hierarchical K-means to make grouping for similar document sources. Figure 3 shows an illustration of clustered document sources in 2 dimensions.

After clustering, we get $C = \{c1, c2, ..., cn\}$ as centroids. Then, transformation of multiple dimensions of centroids is needed to make

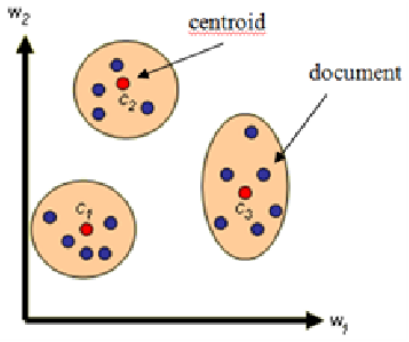

Figure 3. An illustration of clustered document sources in 2 dimensions

ranking and acquire the best clusters. After we get c_i as best centroids, we retrieve all members in c_i as selected document sources for the mentioned classifying words. Assumed $A = \{a1, a2, ..., an\}$ as members in c_i which is selected document sources and $T = \{t1, t2, ..., tn\}$ as collection of words in A as selected aggregate keyword, then we construct the selected metadata as shown in Figure 4.

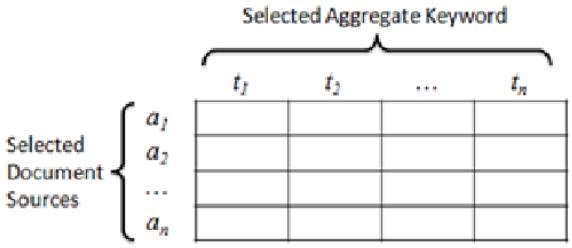

Figure 4. Selected metadata after filtering the source by using a set of classifying words

4.3 Representative Keyword Generation

This section discusses about how to get representative keywords from the selected document sources. First of all, we inverse the selected metadata in Figure 4 in order to promote the selected aggregate keyword to be representative keywords. Then, we apply Hierarchical K-means to make grouping for $T = \{t1, t2, ..., tn\}$ which has high similarity. Figure 5 shows an illustration of clustered selected aggregate keywords in 2 dimensions.

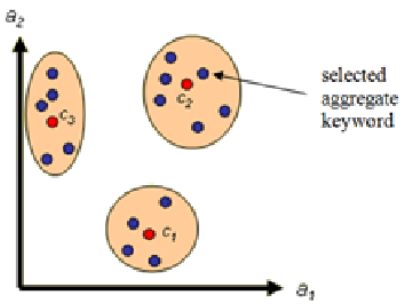

Figure 5. An illustration of clustered selected aggregate keyword in 2 dimensions

After clustering, we get $C = \{c1, c2, ..., cn\}$ as centroids. Then, transformation of multiple dimensions of centroids is needed to rank them and acquire the best clusters. After we get c_i as best centroids, we retrieve all members in c_i as representative keywords. Supposed that $K = \{k1, k2, ..., kn\}$ is members of c_i, and we have several classifying words CW for one field, then the representative keywords are unified K from all CW.

4.4 Metadata Generation

The representative keywords retrieved in previous stage is considered as concept terms. Figure 6 shows the scheme to retrieve the concept terms from one classifying word in case of mud flow porong sidoarjo.

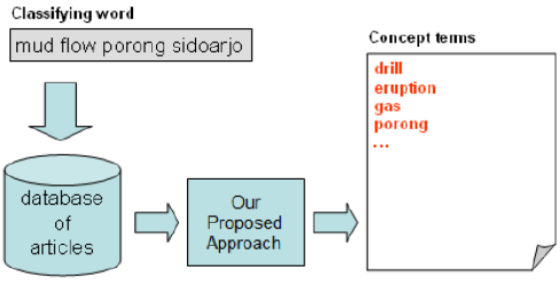

Figure 6. Scheme of concept term generation

In order to generate feature terms of each concept term, we fetch each of concept terms to be a classifying word, as shown in Figure 7.

After acquiring the feature terms of each concept term, the metadata for classifying word

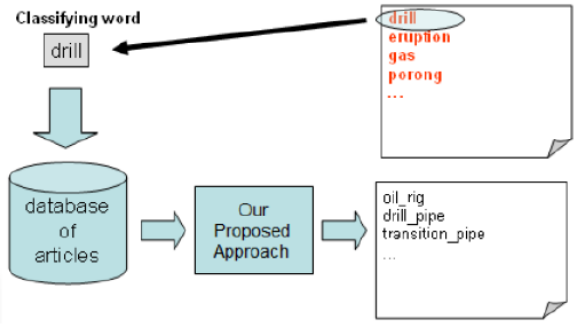

Figure 7. An illustration of clustered selected aggregate keyword in 2 dimensions

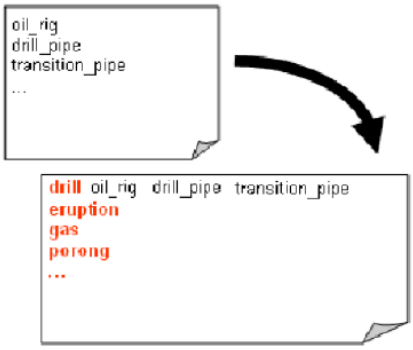

Figure 8. An illustration of metadata generation from concept terms and feature terms

Table 1
Number of article sources and dimensions for each classifying words

Classifying word	Number of Articles	Number of dimensions
Danger	7	1214
Economy	6	846
Environmental	8	1025
Handling	23	1667
Health	2	559
Mud-related	16	1539

"mud flow porong sidoarjo" can be constructed, as shown in Figure 8.

5 Experimental Study

To perform the applicability of our proposed approach for automatic metadata generation,

Ali Ridho Barakbah, "Automatic Metadata Generation by Clustering Extracted Representative Keywords from Heterogeneous Sources",
Industrial Electronics Seminar (IES) 2011, October 26, 2011, Surabaya, Indonesia, and selected for journal publication in the Journal of Emitter, Vol. 2, No. 2, 2012.

Table 2
Metadata of concept terms for each classifying words and ratio of dimension reduction of the keywords

Classifying Word	Number of Extracted Keywords	Concept Terms	Ratio of Dimension Reduction
Danger	26	*area bakrie brantas disaster gas government indonesia java lapindo minister mud people drill environmental flow group million pt business case company cost earthquake president trillion eruption*	97.86%
Economy	31	*bakrie brantas company cost disaster government group lapindo million mud mudflow responsibility sidoarjo day min pay team casing concern drill energi field gas medco operator block case depth feet install negligence*	96.33%
Environmental	45	*aburizal area bakrie brantas call company compensation disaster drill earth environmental expert firm flow gas government group indonesia indonesian java lapindo local minister mud mudflow own people sidoarjo welfare activist controll damage greenpeace responsibility energi international million pt bank campaigner credit dollar fortis friend suisse*	95.6%
Handling	57	*aburizal affect area bakrie begin brantas case company continue cost cubic day disaster drill east fame flow gas government head house indonesian java lapindo million minister month mud mudflow national people porong president problem public resident rp sea sidoarjo team victim village water week work yudhoyono compensation river start stop year indonesia metre pump surabaya volcano*	96.58%
Health	6	*disaster group lapindo mud mudflow control*	98.93%
Mud-related	13	*area company disaster drill flow gas government mud natural bakrie brantas lapindo group*	99.15%

Table 3
Representative keywords of concept terms for all classifying words

Number of Extracted Keywords	Concept Terms	Ratio of Dimension Reduction
97	*aburizal activist affect area bakrie bank begin block brantas business call campaigner case casing company compensation concern continue controll cost credit cubic damage day depth disaster dollar drill earth earthquake east energi environmental eruption expert fame feet field firm flow fortis friend gas government greenpeace group head house indonesia indonesian install international java lapindo local medco metre million min minister month mud mudflow national natural negligence operator own pay people porong president problem pt public pump resident responsibility river rp sea sidoarjo start stop suisse surabaya team trillion victim village volcano water week welfare work year yudhoyono*	96.87%

we conduct an experiment with information sources of Sidoarjo mud flow. We involve 60 English article sources related to Sidoarjo mud flow those have 3095 dimensions of keywords. To filter out the irrelevant sources to get the representative keywords, we set 6 classifying words related to Sidoarjo mud flow; those are Danger, Economy, Environmental, Handling, Health, and Mudrelated. Table 1 shows number of article sources and number of dimensions of mined keywords for each classifying words.

For the technical parameter setting-up of the Hierarchical K-Means that we use for clustering process, we set 10 iteration times of the Hierarchical K-Means using Average Linkage Hierarchical Algorithm with 2 numbers of clusters. We developed our own textmining approach based on the hierarchical structuriza-

Ali Ridho Barakbah, "Automatic Metadata Generation by Clustering Extracted Representative Keywords from Heterogeneous Sources",
Industrial Electronics Seminar (IES) 2011, October 26, 2011, Surabaya, Indonesia, and selected for journal publication in the Journal of Emitter, Vol. 2, No. 2, 2012.

Table 4
Complete metadata consisting of concept terms (indicated by underlined word) and feature terms

aburizal_bakrie indonesia entrepreneuer minister lapindo sidoarjo_hot_mud	limestone sedimentary_rock mineral_calcite calcium_carbonate silica chert flint clay silt sand marine_organism lysocline supersaturated_meteoric_water karst_topography cave acidic_groundwater
basin tectonic strata geological_depression ground_water	medcoenergy energy_company oil gas production drilling_services methanol_production power_generation indonesia bp_migas pertamina sumatra java sulawesi kalimantan natuna
blow_out uncontrolled_release fluid gas petroleum_production	mineral_industry clay natural_gas_liquid sulfur oxigen nitrogen petrol
carbonate_rock sedimentary_rock carbonate_mineral limestone dolomite calcite mineral_dolomite chalk tufa karst_topography cave	mud_volcano volcano magmatic piercement_structure lava_volcano sand_volcano
carbonate mineral salt calcination	oil hydrocarbon transportability strata solvent oil sand liquefied_petroleum_gas
clay_seam aluminium_phyllosilicate_mineral silicate_mineral	parliamentary indonesia representation people
drill_string oil_rig drill_pipe kelly_drive top_drive drill_collars drilling_fluid column string annulus ring_shaped_void bottomhole_assembly transition_pipe drill_bit	pertamina natural_resource oil_industry expert oil_academy gas_mining refinement drilling
earthquake earth crust seismic_wave seismometer seismograph magnitude richter_scale damage tsunami destruction tectonic_plate hypocenter epicenter	porong subdistrict sidoarjo east_java indonesia
east_java province java surabaya indonesia	power_transmission energy power transmission
fish_pond water fish pond embankment lake	railway_track railway steel_rail cross_ties tramway_track light_rail drainage heavy_rail tramway_track
gas natural_gas hydrocarbon hydrogen sulphide nitrogen_dioxide ethyl_mercaptane methyl_mercaptane caolin sulfur	rice_paddy arable_land rice semiaquatic_crop irrigation
gas_pipeline oil natural_gas rail_road tanker steel plastic pump crude_oil paraffin pipeline hazardous material corrosion	santos oil gas cooper_basin natural_gas_processing oil_and_gas_exploration indonesia
government authority power rule law people organization control	shrimp_pond prawn embankment disease production export industry monoculture_industry population ecology
graben depressed valley horst dip	sidoarjo regency gerbangkertosusila east_java indonesia
hidrogen_sulphide h2s colorless toxic flammable_gas rotten_egg bacterial swamp sewer anaerobic_disgestion volcanic_gas natural_gas well_water archaic	steel_casing encasement_pipe underground_construction underground_boring pipe_connector drill watermain natural_gas electrical_high_voltage_line optic_communication_line casing_end_seal
indonesia islands archipelagic jakarta bhinneka_tunggal_ika pancasila	surabaya east_java indonesia
java_island java indonesia bengawan_solo_river borobudur wali_songo mount_bromo	susilo_bambang_yudhoyono president indonesia election pacitan east_java
jusuf_kalla indonesia vice_president golkar south_sulawesi	toll_road toll_way pike tollpike toll_booth toll_plaza toll_gate
kujung_formation limestone drilling_layer sediment_stone natural_gas reservoir_gas	village community rural_area
lapindo_brantas oil gas bakrie_group aburizal_bakrie lapindo_brantas sidoarjo_mud_flow	volcanic_debris volcanic_ash rock mineral volcanic_vent solid_rock magma volcanic_activity ash steam tephra pyroclastic_debris cinder
	walhi environment indonesia organization

tion of English grammar in order to extract the keywords from all sources.

After we used a set of the classifying words, we apply our approach described in previous section to generate the metadata of concept terms. Table 2 shows the metadata of concept terms for each classifying words and ratio of dimension reduction of the keywords. After the representative keywords are extracted for each classifying words, we generate the representative keywords for all classifying words and ratio of dimension reduction of the keywords as shown in Table 3. Table 3 performs that our proposed approach is able to reduce 96.87% of the metadata space from 3095 to 97 dimensions.

To generate the feature terms of representative concept terms, each representative concept term can be supposed as a classifying word, and then it will extract a series of representative keywords as feature terms.

Our proposed approach for generating the metadata of heterogeneous sources also can be used with the manual selection of concept terms by human. In this case, the selected concept terms are supposed as classifying words and then it will extract the representative feature terms. Table 4 shows the complete metadata consisting of concept terms (indicated by underlined word) (by human selection) and feature terms.

6 CONCLUSION

In this paper we presented a new approach to automatically generate a representative metadata by applying a clustering in order to extract representative keywords from heterogeneous sources. The proposed approach consists of three stages: (1) Aggregate Keyword Extraction, (2) Automatic Source Filter, and (3) Representative Keyword Generation. To perform the applicability of our proposed approach for automatic metadata generation, we conducted an experiment with information sources of Sidoarjo mud flow consisting of 60 English article sources related to Sidoarjo mud flow. we set 6 classifying words related to Sidoarjo mud flow; those are Danger, Economy, Environmental, Handling, Health, and Mud-related. Our proposed approach then extracted the representative keywords and is able to reduce the dimensions of keywords for each classifying words. Finally, our proposed approach can generate the representative metadata and is able to reduce 96.87% of the metadata space from 3095 to 97 dimensions. The experimental result performed effectiveness of the proposed approach to generate the representative metadata and reduce drastically the metadata space of the keywords.

ACKNOWLEDGEMENT

We greatly appreciate the help provided by Dr. Dadet Pramadihanto, Prof. Yasushi Kiyoki, Dr. Takafumi Nakanishi, all members of Knowledge Engineering of Global Disaster Risk Management in Electronic Engineering Polytechnic Institute of Surabaya (EEPIS) Indonesia, and Knowledge Clustered Group in National Institute of Information and Communications (NICT) Japan.

REFERENCES

[1] D. Sakai, Y. Kiyoki, N. Yoshida, and T. Kitagawa, A Semantic Information Filtering and Clustering Method for Document Data with a Context Recognition Mechanism, Information Modelling and Knowledge Bases (IOS Press), Vol. XIII, May 2003.

[2] Kohei Arai and Ali Ridho Barakbah, Hierarchical Kmeans: an algorithm for centroids initialization for K-means, Reports of the Faculty of Science and Engineering, Vol. 36, No. 1, 2007.

[3] P. Pathak, M.D. Gordon, and W. Fan, "Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation", Proceedings of the 1999 Americas Conference on Information Systems August 13-15, 1999, Milwaukee, WI, USA.

[4] S. Sasaki, Y. Kiyoki, and H. Akutsu, An application of Semantic Information Retrieval System for International Relations, Information Modelling and Knowledge Bases (IOS Press), Vol. XVIII, May 2007.

[5] W. Fan, M.D. Gordon, and P. Pathak, "Discovery of Context-Specific Ranking Functions for Effective Information Retrieval Using Genetic Programming", IEEE Transactions on Knowledge and Data Engineering, 16(4), 523-527, 2004.

[6] W. Fan, M.D. Gordon, P. Pathak, W. Xi, and E.A. Fox, "Ranking Function Optimization For Effective Web Search By Genetic Programming: An Empirical Study", Proceedings of the 33rd Hawaii International Conference on System Science (HICSS), 2000, Hawaii, USA.

[7] Y. Kiyoki and S. Ishihara, Semantic search space integration method for meta-level knowledge acquisition from heterogeneous databases, Information Modelling and Knowledge Bases (IOS Press), Vol. XIV, pp. 86-103, May 2002.

Ali Ridho Barakbah received his Bachelor degree from Department of Informatics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia in 1997 and PhD degree from Graduate School of Media and Governance, Keio University, Japan in 2011. He works with Electrical Engineering Politechnic Institute of Surabaya. His research interests are in the field of Intelligent Computing and Knowledge Engineering.