Plagiarism Checker X Originality Report

Similarity Found: 15%

Date: Sunday, December 22, 2019
Statistics: 299 words Plagiarized / 2047 Total words
Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE The Development of Document Similarity Detector by Jaccard Formulation Darwis Robinson Manalu, Edward Rajagukguk, Rimbun Siringoringo Fakultas Ilmu Komputer, Universitas Methodist Indonesia Medan, Indonesia manaludarwis@gmail.com Desmon Kristanto Siahaan Departmen of Computer Science and Information Engineering, National Dong Hwa University Taiwan desmonsiahaan02@gmail.com Poltak Sihombing Departemen Ilmu Komputer Universitas Sumatera Utara Medan, Indonesia poltaksihombing@yahoo.com Abstract In this research, we proposed the development of the Document Similarity Detector (DSD) by Jaccard formulation.

By this formulation, we can calculate the document similarity based on fitness value. In the retrieval process, we modified the keywords of documents to be the chromosome in the Genetic Algorithm (GA) process in finding the Keyword Solution (KS) by Keyword Competition. Keyword Competition means to eliminate some of keywords based on fitness value of Jaccard.

As the data testing, we used the collection paper in Proceeding of SNIKOM, in Medan. We used the GA method to eliminate some of keywords in order to get the most potential keywords of documents in database. We then calculate the appearance of keywords that resulted by GA in the paper collection of documents in the database.

This research resulted the similarity value of documents retrieved from database. Keywords — Jaccard formulation, Genetic Algorithm, keyword competition, document similarity, retrieval. I. INTRODUCTION Nowadays, the increasing of documents in the database growth very fast.

The challenge is more on the retrieval of the right documents since documents stored in a database grow very fast and soon become unmanageable. This situation often resulted in difficulty to retrieve a document from a database which is expected to be very relevant to a [1]. The goal of any Retrieval System is to help a user to locate the right document or those documents that have the potential to satisfy his information need.

To solve this problem, researchers have implemented some methods such as inverted index, Boolean querying, knowledge- based, neural network, probabilistic retrieval and machine learning approach, etc. The most frequently used techniques include symbolic, inductive learning algorithms, multiple- layered, feed-forward neural networks such as back propagation networks, and evolution-based genetic algorithm (GA) [6].

The main problem in this paper is a discussion about the document similarity testing using Jaccard function in genetic algorithm process. In this research, we used the paper collection of SNIKOM (Seminar Nasional Ilmu Komputer) as the datatest. The GA was used because of its ability to optimizing of document retrieval by keywords competition.

In GA process a chromosome is generated by gene which represents bit (0 and 1)[5]. We develop the prototype of document similarity detector of document retrieval by Jaccard formulation called it DSD (Document Similarity Detector) [12]. By this prototype we can test how many similar a document to the other one in a database II. METHODOLOGY A.

The Typical of System Figure 1 shows the methodology of a typical our development systems. A document collection is represented by some of keywords. In the first step, those keywords will be modified to be chromosomes[8]. All of chromosomes will be selected based on fitness value by Jaccard formulation in GA process.

A user who wishes to search this document collection expresses an information need in the form of query that is posed to the system. By this experimental set up, we can see the results of GA scheme document retrieval similarity by Jaccard similarity formulation[9]. The user can interacted with this ranked list by indicating documents that related to his information need. B.

Jaccard formulation We proposed the document similarity detector based on fitness value by Jaccard fitness formulation[10]. To be 1 in document 1 = k1 x d1, and the appearance of k2 in document 2 = k2 x d2, and the appearance of KS ki in document j = ki x dj and the appearance of total KS ki,...n in all of documents collection = Ktotal. The Jaccards formulation as shown in equation 1 as the common measure of term or keyword[2][3][4].

In this way, the X & Y symbols represent the element of document. If #(X) is equal to #(Y) then Jaccard value is 1, (the symbol # means number of element X, X means element of document X and Y means element of document Y). That means if X element that document X is precisely the same with document Y. Element of document may be represented by terms or keywords.

If a keyword is presented in a document, the bit is set to 1; otherwise it is set to 0. Each document could be represented in terms of a sequence of 0s and 1s. Document with a higher Jaccard higher probability of similarity. Jaccard formulation is shown as following: = (1) Where #(S) showing number of element in S, for example: S = {a, b, c, d, e, f, g, h, i, j}; If X = {a, b, e, g, h, i, j}; and Y = {b, c, d, f, g, j} Figure 1: The Typical of System III. PROTOTYPE We have developed the Prototype of DSD formulation.

In the prototype, the DSD formulation will determine the similarity measure of document retrieved from database. The DSD prototype is shown in figure 2. Figure 2. The Prototype of DSD We have developed a Document Similarity Detector in DSD prototype as is shown in figure 2. In the prototype, keyword Competition Approach by fitness value in Jaccard formulation will determine which the chromosome is having the opportunity to be continued in the next generation. Fitness is a value that is used to select the chromosome for the next generation process.

The objective of this phase is to obtain fitness value of chromosome. Fitness of chromosome is a value that used in selecting each chromosome for the next generation. Fitness value evaluation depends on the case of each chromosome in each generation. We have implemented the Jaccrs fitness functions in this research as shown in the DSD prototype as is shown in figure 2. IV.

RESULTED The last process in GA process will result the keywords solution as the last keyword. The last keyword called keyword solution or KS will be linked to documents collection in database. In the next process, the system will count how many times the appearance of KS and rank document retrieved according to each KS.

The DSD prototype then present the document retrieved as shown in table 1. TABLE 1. THE DSD RESULTED Rank Doc. Similarity (%) Paper- Id The Title of Documents collection 1 36.11% 165 Study of Compression Data Arithmetic Coding and Cryptography RSA 2 23.21% 274 Securing text messages with a combination of one-time pad cryptographic algorithms and playfair cipher 3 20.67% 542 Type of Attack on Digital Image Watermarking Spread Spectrum 4 14.31% 172 Increasing Digital Audio Watermarking Robustness Through MSB and Algortima RSA 5 12.38% 189 Recognition of Digital

Signatures Using the Learning Vector Quantization (LVQ) Method Table 1 presents the example of similarity calculation of paper collection retrieved. We find out that the similarity percentage is: 36.11% in paper-Id 165, followed by paper-Id 274 (23.21%), paper-Id 542 (20.67%), paper-Id 172 (14.31%), and paper-Id 189 (12.38%). V. CONCLUSION AND FUTURE WORK A.

Conclusion In this research we have proposed proposed the Jaccards formulation as the fitness in GA Process in order to get the keyword solution in Keyword Competition (KC) scheme. The KC scheme have been tested in the database of paper collection of SNIKOM (Seminar Nasional Ilmu Komputer) in Medan. Information of Document; Id;title;keyword; year;Abstract.

ferensi Clustering Data test Query GA Keywords Jaccard Result Users Searching By looking into the relationship of number of queries, we can conclude that if the query increases, the similarity value does not increased. This supports our propose work which proposed that to get the potential relevant document (more similar) from database, we need to search by the potential keyword, and to do this it is important to select the set of asers kerosed in tesech [9].

The behavior of GA process on similarity in KC scheme, we have observed that if the number of query increases, therefore the sum of generation, crossover, mutation and process time also increased. We conclude that if the number of queries, crossover, and mutation increases, then the similarity percentage of document retrieved does not increase.

From the effect of sum of generation in KC scheme, we can conclude that even though the sum of generation increases, the percentage of similarity value does not necessarily increase. In terms of process time generation, we conclude that if the sum of generation increases, the time of process also increases. B. Future work The work presented in this research is by no means complete.

Rather, we believe that the ideas are a starting point for a number of interesting projects, either related to information retrieval. We have tested in this research, KC approach and keyword based ranking scheme have succeeded in implementation in DSD prototype. We hope this prototype can be developed further to the large scope such as plagiarism detector.

In a future work, we outline some particular projects which we think could be rewarding. These projects are; to the improvement of information access systems in networks and build it in the digital library by multi dimensional searching in the larger database.

ACKNOWLEDGMENT The authors would like to thank the Dean Fakultas Ilmu Komputer Universitas Methodist Indonesia (UMI) in Medan for the excellent services to support this research, Tnd stf “Lega Plitn da engabdian pada Masyarakat (LP3M), To the Rector, Universitas Methodist Indonesia, Medan, for their widest to support this work. VI. REFERENCES [1] D. Soyusiawaty and Y. Zakaria, "Book Data Content Similarity Detector With Cosine Similarity (Case study on digilib.uad.ac.id)," International Conference on Telecommunication Systems, Services, and Applications (TSSA), 2018. [2] D.

Soyusiawaty, "Designing and Implementing Parsing for Ambiguous Sentences in Indonesian Language," Journal of Theoretical and Applied Information Technology, vol. 84, no. 3, pp. 339-347, 2016. [3] S. Forman and B. K. Samanthula, "Secure Similar Document Detection: Optimized Computation Using the Jaccard Coefficient," in IEEE, 2018. [4] P. K. Verma, S. Agarwal and M. A.

Khan, "Opinion Mining considering Roman Words using Jaccard Similarity Algorithm based on Clustering," in ICCCNT, Delhi, India, 2017. [5] R. S. N. and M. D. A., "Jaccard index based clustering algorithm for mining online review," nternational Journal of Computer Applications,, vol. 105, no. 15, 2014. [6] E. B. Nababan, O. S. Sitompul and C.

Yuni, "Genetic Algorithms Dynamic Population Size with Cloning in Solving Traveling Salesman Problem," Data Science: Journal of Computing and Applied Informatics (JoCAI), vol. 2, no. 2, 2018. [7] Anadakumar and P. , "A Survey on Preprocessing in Text Mining," International Journal of Advanced Research in Computer Science, vol. 9, no. 4, p. 79 – 91, 2013. [8] T. Kato, I.

Shimizu and T. Pajdla , "Selecting image pairs for SfM by introducing Jaccard Similarity," IPSJ Transactions on Computer Vision and Applications, vol. 9, no. 12, 2017. [9] J. Lee and R. Tukhvatov, "Evaluations of Similarity Measures on VK for Link Prediction," Data Science and Engineering, vol. 3, no. 3, p. 277 – 289, 2018. [10] S. Niwattanakul and J.

Singthongchai, , "Using of Jaccard Coefficient for Keywords Similarity," in Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hongkong, 2013. [11] S. Sugiyamto, B. Surarso and A. Sugiharto, "Analisa Ferforma Metoda Cosine dan Jaccard Pada Pengujian Kesamaan Dokumen," Jurnal Masyarakat Informatika, vol. 5, no. 10, 2014. [12] D. R.

Manalu and E. Rajagukguk, "Pengujian Tingkat Kemiripan Skripsi Mahasiswa Degnan Algoritma Genetika Menggunakan POSI Formulation," Methodika, vol. 2, no. 2, pp. 175-182, 2016. [13] D. R. Manalu and D. R. Sitompul, "Developing Learning Media of

Teaching of 8051 Microprocessor in Data Retrieval to Support the ALFHE," Internet, vol. 8, no. 1, 2017. [14] P.

Sihombing, "Kompetisi Keyword Pada Algoritma Genetika dengan Fitnes Jaccard dan Dice," in SeNAIK, Indonesia, 2013. The Author 1. (Darwis Robinson Manalu), as Lecturer of Fakultas Ilmu Komputer, Universitas Methodist Indonesia (UMI), Medan. 2. (Poltak Sihombing), as the head of Computer Science Department; Faculty of Computer Science and Information Technology (Fasilkom-TI); Universitas Sumatera Utara ; Medan; Indonesia. 3. (Edward Rajagukguk), as Lecturer of Fakultas Ilmu Komputer, Universitas Methodist Indonesia (UMI), Medan 4.

(Rimbun Siringoringo), as Lecturer of Fakultas Ilmu Komputer, Universitas Methodist Indonesia, Medan 5. Desmon Kristanto Siahaan, as Student of Department of Computer Science and Information Engineering, National Dong Hwa University, Taiwan

INTERNET SOURCES:

1% - https://www.facebook.com/Fakultas-Ilmu-Komputer-Universitas-Methodist-Indonesia-226420824073845/
<1% - http://repository.usu.ac.id/bitstream/handle/123456789/53366/Cover.pdf;sequence=7
<1% - https://amds123.github.io/page198/
<1% - https://www.researchgate.net/post/How_can_we_get_the_similarity_score_between_two_vectors
2% - https://core.ac.uk/download/pdf/11959824.pdf
<1% - https://biodatamining.biomedcentral.com/articles/10.1186/s13040-015-0061-5
1% - https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/chen27.pdf
1% - https://globaljournals.org/GJCST_Volume15/1-QFSRD-Orthogenesis-Evolution.pdf
<1% - https://repository.maranatha.edu/view/year/2015.type.html
<1% - https://www.sciencedirect.com/book/9780124438705/fuzzy-theory-systems
<1% - https://www.researchgate.net/publication/3322044_An_introduction_to_voice_search
<1% - http://ijarcsse.com/Before_August_2017/docs/papers/Volume_3/8_August2013/V3I8-0303.pdf
<1% - http://www.umiacs.umd.edu/~jimmylin/publications/Efron_etal_SIGIR2014.pdf
<1% -

https://www.researchgate.net/publication/267838711_Sonification_of_an_information_retrieval_environment_Design_issues

<1% - https://docs.google.com/document/d/1dhmqzHI2b-CfDwPk5506xW5zB4VHhyF8UaWPlQm_GpU/preview#!

<1% - http://shodhganga.inflibnet.ac.in/bitstream/10603/27879/10/10_chapter%204.pdf

<1% - http://shodhganga.inflibnet.ac.in/bitstream/10603/32680/16/16_chapter%206.pdf

<1% - https://www.staff.ncl.ac.uk/damian.giaouris/pdf/Papers/Tuning%20of%20PI%20Speed%20Controller%20in%20DTC%20of%20Induction%20Motor%20Based%20on%20Genetic%20Algorithms%20and%20Fuzzy%20Logic%20Schemes.pdf

<1% - http://www.ece.uprm.edu/~mvelez/research/tcess/IPEG%202006.doc

<1% - http://www.scirp.org/journal/paperinformation.aspx?paperid=95498

<1% - https://www.sciencedirect.com/science/article/pii/S0020025519308588

<1% - https://beyondintractability.org/mbi-cci-fall-update

1% - http://tssa-conference.org/2017/wp-content/uploads/2017/10/Technical-Program-TSSA-2017-v1.1.pdf

1% - http://www.jatit.org/volumes/Vol84No3/5Vol84No3.pdf

1% - http://garuda.ristekdikti.go.id/journal/view/13828?page=2

<1% - https://www.researchgate.net/publication/310500305_Opinion_mining_and_fuzzy_quantification_in_hotel_reviews

1% - https://nrid.nii.ac.jp/nrid/1000070312915/

<1% - https://link.springer.com/content/pdf/10.1007%2Fs41019-018-0073-5.pdf

1% - https://link.springer.com/chapter/10.1007/978-3-319-60240-0_35

<1% - http://scholar.google.co.id/citations?user=-Q7OSsQAAAAJ&hl=en

1% - http://www.methodist.ac.id:8082/cdn/File/Darwis%20Reviewer%20Jurnal/Plagiat/Alfhe%20Checker.pdf

<1% - https://jurnal.wicida.ac.id/index.php/senaik/issue/view/20

<1% - http://repository.usu.ac.id/bitstream/handle/123456789/55936/Cover.pdf?sequence=6&isAllowed=y

<1% - https://iopscience.iop.org/issue/1742-6596/1235/1

<1% - http://naikson.org/?page=alumni

<1% - http://scholar.google.co.id/citations?user=-Wht5rYAAAAJ&hl=en