Analisis Frekuensi Kata untuk Mengekstrak Kata Kunci dari Artikel Ilmiah Berbahasa Indonesia

Ni Wayan Sri Arini[1], Ida Bagus Putu Widja[2], Komang Rinartha[3]
Program Studi Sistem Komputer, STMIK STIKOM Bali
Jalan Raya Puputan Renon no. 86 Denpasar, 0361-244445
e-mail: [1]baliaditya13@gmail.com , [2]ibpwidja@gmail.com , [3]komangrinartha@gmail.com

Abstrak

Publikasi hasil penelitian merupakan suatu proses yang harus dilaksanakan dalam sebuah kegiatan penelitian. Publikasi dapat dilaksanakan dalam bentuk presentasi dalam sebuah seminar ilmiah, maupun dalam bentuk jurnal ilmiah. Sebelum memasuki proses seleksi, artikel ilmiah tersebut dipilah sesuai dengan kompetensi yang dimiliki oleh tim penilai. Umumnya proses pemilahan artikel ilmiah dilakukan secara manual oleh panitia pengelola seminar ilmiah, sehingga membutuhkan waktu dan membutuhkan ketepatan dalam penentuan tim penilai yang sesuai dengan artikel ilmiah. Pemilahan artikel ilmiah dapat dilakukan dengan menerapkan algoritma string similarity, yaitu dengan mencari kata-kata kunci yang terdapat dalam sebuah karya ilmiah. Kata kunci yang berada dalam artikel yang dihasilkan berdasarkan frekuensi kata yang muncul. Sebelum dicari kata yang banyak muncul, dilakukan proses filtering untuk menghilangkan kata sambung yang sering muncul sehingga tidak dianggap sebagai kata kunci artikel. Filtering menggunakan data stopword list yang digunakan oleh Tala. Sistem dibangun dalam bentuk aplikasi web menggunakan bahasa pemrograman PHP dan database MySQL dengan teknik responsive web design. Hasil penelitian ini menjelaskan bahwa artikel yang dimasukkan ke dalam sistem dapat dihasilkan kembali kata kunci yang sesuai dengan mendata kata-kata yang banyak muncul.

Kata kunci: *Frekuensi Kata, Artikel Ilmiah, Kata Kunci.*

Abstract

Publication of research results are a process that must be carried out in a research activity. Publications can be carried out in the form of presentations at a scientific seminar, as well as in the form of scientific journals. Before entering the selection process, the scientific article is sorted according to the appropriate reviewer's competence. Generally, the process of sorting scientific articles is done manually by the scientific seminar management committee, so it takes time and requires accuracy in determining reviewer that is appropriate in scientific articles. Sorting scientific articles can be done by applying a string similarity algorithm, which is by searching for key words contained in a scientific article. The keywords in the article are generated based on the frequency of the words that appear. Before searching for a word that appears a lot, a filtering process is carried out to eliminate the conjunctions that often appear so that they are not considered as keyword articles. Filtering process using Tala's Stopword list. The system is built in the form of web applications using the PHP programming language and MySQL database with responsive web design techniques. The results of this research describe that articles have been entered into the system can be traced back to keywords by recording the words that appear a lot.

Keywords*: Word Frequency, Scientific Article, Keywords.*

1. Pendahuluan

Publikasi hasil penelitian harus dilaksanakan oleh kalangan peneliti, dosen, maupun para mahasiswa yang telah melaksanakan kegiatan penelitian. Publikasi hasil penelitian juga sedang dikembangkan di Indonesia agar terindeks secara nasional, tidak hanya terindeks secara internasional. Publikasi hasil penelitian ini bertujuan untuk menyebarluaskan hasil penelitian di kalangan masyarakat umum dan juga di kalangan masyarakat peneliti dan akademisi. Publikasi hasil penelitian dapat dilaksanakan dalam bentuk seminar ilmiah baik skala nasional maupun internasional. Selain itu juga dapat berupa jurnal nasional maupun jurnal internasional.

Publikasi dalam bentuk jurnal umumnya tidak terlalu banyak artikel yang dimuat di dalamnya, namun beda halnya dengan seminar yang umumnya memuat banyak artikel ilmiah di dalamnya. Karena

banyak artikel yang akan dimuat dalam sebuah seminar ilmiah, maka diperlukan banyak *reviewer* dengan berbagai bidang keahlian yang dimilikinya. Ketika artikel ilmiah diterima oleh pengelola seminar ilmiah, maka artikel akan dipilah untuk disampaikan kepada para *reviewer* sesuai dengan bidang ilmu yang dimiliki. Proses pemilahan ini tidak dapat dilakukan dengan waktu yang sangat singkat dan pemilihan *reviewer* yang tepat sesuai dengan topik penelitian yang dimuat. Panitia pengelola seminar harus membaca artikel terlebih dahulu agar mendapatkan *reviewer* yang sesuai.

Algoritma TF-IDF merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap *token* (kata) di setiap dokumen dalam *korpus* [1]. Pada *term frequency* akan menghitung jumlah kata yang muncul dalam sebuah dokumen, sehingga dapat digunakan untuk mencari kata kunci dalam dokumen. Pencarian kata kunci tidak hanya dilakukan pada abstrak saja, namun secara keseluruhan dari artikel ilmiah. Mikhail Alexandrov dkk, telah mencoba mengelompokkan artikel ilmiah berdasarkan abstrak hasilnya tidak sebaik mengelompokkan artikel ilmiah secara keseluruhan [2].

Ekstraksi kata kunci atau *keyword extraction* didefinisikan sebagai kegiatan yang secara otomatis mengidentifikasi sekumpulan istilah yang paling tepat menggambarkan subjek dokumen. Terminologi yang berbeda digunakan dalam mempelajari istilah-istilah yang mewakili informasi paling relevan yang terdapat dalam dokumen: frasa kunci, segmen kunci, istilah utama atau hanya kata kunci. Semua sinonim yang terdaftar memiliki fungsi yang sama mencirikan topik yang dibahas dalam dokumen. Mengekstrak sekumpulan unit kecil yang terdiri dari satu atau lebih istilah dari satu dokumen merupakan masalah penting dalam *Text Mining* (TM), *Information Retrieval* (IR) dan *Natural Language Processing* (NLP). Kata kunci banyak digunakan untuk mengaktifkan *query* dalam sistem IR karena mudah didefinisikan, direvisi, diingat, dan dibagikan. Dengan kata lain, kata kunci yang diekstrak dapat digunakan untuk membuat indeks otomatis pada pengumpulan dokumen atau dapat digunakan untuk representasi dokumen dalam kegiatan kategorisasi atau klasifikasi. Sementara menetapkan kata kunci ke dokumen secara manual sangat memakan waktu dan kegiatan yang membosankan dan sebagai tambahan, jumlah dokumen yang tersedia secara digital semakin berkembang [3].

Algoritma *string similarity* dapat digunakan untuk menghitung tingkat kesamaan beberapa dokumen berdasarkan kata yang ada dalam dokumen-dokumen tersebut. *String similarity* dapat digunakan untuk membandingkan antara kata kunci dari sebuah artikel ilmiah dengan bidang keahlian yang dimiliki oleh *reviewer*. Beberapa metode *string similarity* dapat digunakan untuk membandingkan kata kunci artikel ilmiah dengan bidang keahlian *reviewer* salah satunya adalah *Cosine Similarity*.

Website merupakan salah satu teknologi yang umum digunakan saat ini dalam penerimaan artikel ilmiah. Teknologi *web* yang banyak berkembang saat ini adalah teknologi *web responsive*. *Web responsive* adalah desain *web* yang dapat digunakan secara fleksibel dibuka dari komputer dengan berbagai resolusi layar. Selain dengan teknologi *responsive*, *website* juga dibangun dengan menggunakan bahasa pemrograman PHP serta *database* MySQL. Web konvensional pada umumnya tidak dapat dilihat dengan baik pada resolusi *display* yang berbeda-beda.

Maryam Habibi dan Andrei Popescu-Belis dalam penelitiannya menyatakan bahwa banyak metode telah diusulkan untuk mengekstrak kata kunci dari teks secara otomatis, dan berlaku juga untuk percakapan yang ditranskripsikan. Teknik paling awal menggunakan frekuensi kata dan nilai TF-IDF untuk menentukan peringkat kata untuk ekstraksi [4].

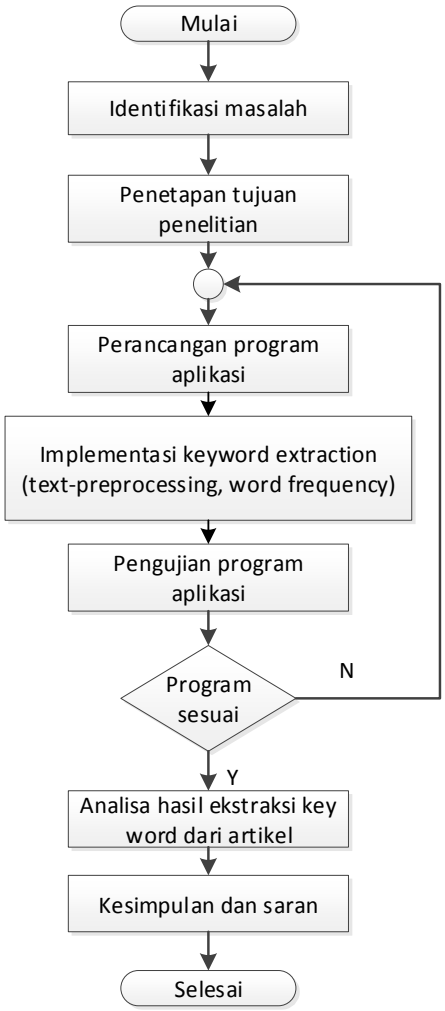
Utami Dewi dan Imelda, dalam penelitiannya menjelaskan bahwa *text mining* merupakan proses pengambilan data berupa teks dari sebuah sumber dalam hal ini sumbernya adalah dokumen. Dengan *text mining* dapat dicari kata-kata kunci yang dapat mewakili isi dari suatu dokumen lalu dianalisis dan dilakukan pencocokan antara dokumen dengan *database* kata kunci yang telah dibuat untuk menentukan atau memilah kategori suatu dokumen [5].

Abdul Azis Maarif dalam penelitiannya menjelaskan bahwa memilah artikel ilmiah bisa dilakukan dengan mudah oleh manusia, namun pemilahan dokumen yang dilakukan secara otomatis oleh komputer akan membawa masalah tersendiri. Begitu pula dengan mengukur tingkat kesamaan dokumen dengan dokumen lain, manusia dapat dengan mudah mengukur apakah suatu dokumen memiliki tingkat kesamaan dengan dokumen lainnya. Kata kunci yang digunakan dalam proses ekstraksi dokumen dalam proses pengurutan kategori dokumen. Agar hasil pengukuran tingkat kesamaan dokumen dengan kata kunci mendapatkan hasil yang optimal maka algoritma yang digunakan untuk algoritma *text mining* yang digunakan dalam proses dimana TF-IDF (*Term Frekuensi - Inversed Document Frequency*) dari model IR (*information retrieval*) sebagai ukuran tingkat kesamaan antara dokumen dengan kata kunci yang diperoleh dari ekstraksi teks dalam dokumen [1].

Mikhail Alexandrov, dkk dalam penelitiannya menjelaskan bahwa akses gratis ke artikel ilmiah lengkap di perpustakaan digital utama dan *repository web* lainnya terbatas hanya pada abstrak. Mikhail Alexandrov, dkk menggunakan metode Stein's MajorClust untuk mengelompokkan kata kunci dan dokumen. Percobaan pendahuluan Mikhail Alexandrov, dkk menunjukkan bahwa abstrak tidak dapat dikelompokkan dengan kualitas yang sama seperti artikel lengkap, meskipun kualitas yang dicapai memadai untuk banyak aplikasi [2].

2. Metode Penelitian

Untuk mempermudah proses penelitian, maka diperlukan adanya perencanaan penelitian dari persiapan penelitian hingga pengambilan kesimpulan dan saran. Dari tahap awal penelitian, penelitian dilakukan dengan identifikasi permasalahan yang ada yang kemudian dalam hal ini adalah mengetahui konsep *keyword extraction* dari sebuah artikel ilmiah dan penentuan *reviewer* artikel ilmiah berdasarkan kata kunci yang didapatkan. Setelah mengetahui tujuan penelitian, langkah selanjutnya adalah melakukan analisa *word frequency* dan *text pre-processing*. Setelah melakukan analisa teori dari metode yang digunakan, langkah selanjutnya adalah melakukan perancangan terhadap aplikasi yang akan dibangun. Setelah perancangan selesai, maka dilakukan implementasi perangkat lunak dengan menggunakan bahasa pemrograman PHP dan *database* MySQL dalam bentuk *responsive web design*. Setelah implementasi dari program dengan metode tersebut, maka dilakukan pengujian terhadap metode yang digunakan, apabila program tidak sesuai dengan analisa teori yang ada, maka kembali ke proses perancangan, jika program sudah sesuai dengan teori yang ada, maka dilanjutkan dengan analisa hasil dari ekstrak kata kunci. Setelah mendapatkan hasil dari analisa ekstraksi kata kunci, maka diambil kesimpulan atas penelitian yang telah dibuat. Adapun gambar tahapan penelitian ditunjukkan pada Gambar 1.

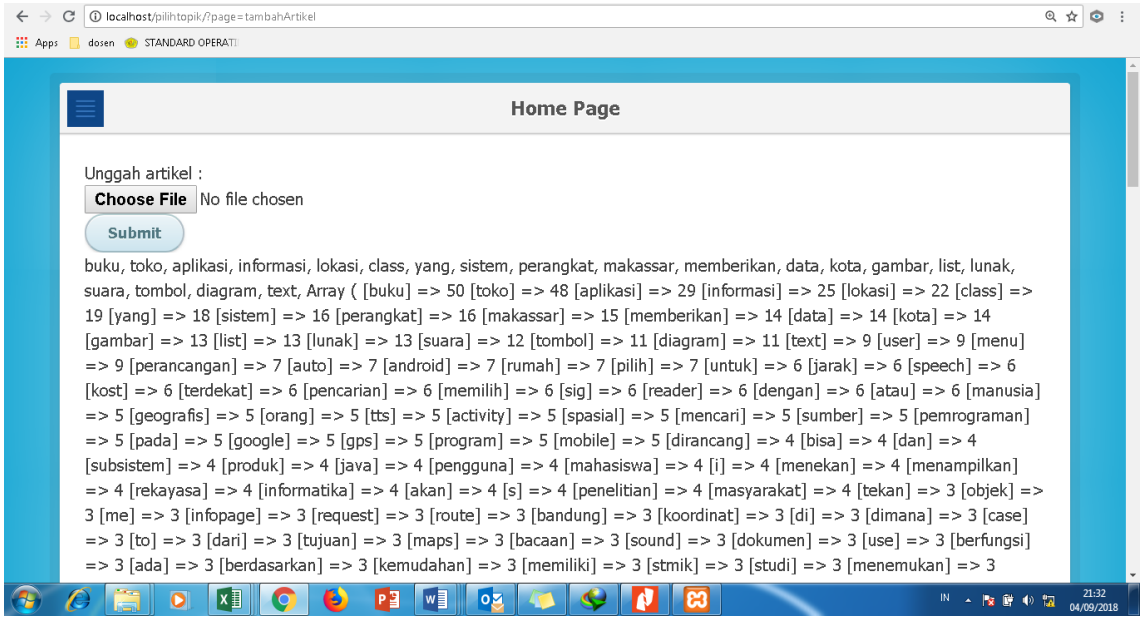


Gambar 1. Tahapan penelitian.

3. Hasil dan Analisis

Proses yang dilakukan adalah melakukan ekstraksi kata yang berada di dalam sebuah artikel menggunakan frekuensi kata. Sebelum menghitung frekuensi kata yang terdapat dalam artikel, terlebih dahulu artikel melalui proses *filtering* yaitu menghilangkan kata-kata penghubung yang sering terdapat dalam sebuah kalimat. Proses *filtering* ini menggunakan data *stopword* yang digunakan dalam penelitian Fadillah Z Tala.

Dari percobaan didapatkan bahwa aplikasi dapat menghasilkan kata yang mewakili artikel yang dimasukkan. Kata kunci yang dihasilkan dari artikel diambil hanya 20 kata dengan frekuensi penggunaan tertinggi dengan asumsi 20 kata tersebut dapat mewakili artikel yang dimasukkan. Percobaan dilakukan menggunakan perangkat lunak berbasis web dengan bahasa pemrograman PHP menggunakan teknologi *responsive web design*. Adapun tampilan modul pengujian sistem ditunjukkan pada Gambar 2.



Gambar 2. Tampilan modul pengujian *word frequency*.

Tabel 1. Beberapa contoh ekstraksi kata menggunakan frekuensi kata.

Judul	Kata kunci artikel	Kata kunci frekuensi
Penggolongan Musik Terhadap Suasana Hati Menggunakan Metode K-Means	Musik, mood, pengelompokkan, clustering, k-means	data, musik, cluster, mood, nilai, clustering, means, sistem, yang, file, proses, penelitian, rata, untuk, hati, tahapan, suasana, algoritma, bagian, memiliki
Penerapan Metode Analytical Hierarky Process (AHP) Pada Penentuan Pemberian Kredit Usaha Rakyat Berbasis Web Pada PT.Bank Rakyat Indonesia	KUR, Analytical Hierarcy Process (AHP), website	nasabah, kredit, gambar, kriteria, activity, diagram, metode, data, sistem, bank, usaha, pengajuan, admin, penelitian, rakyat, keputusan, input, analytical, perbandingan, process
Aplikasi Sistem Informasi Persediaan Barang Pada PT Sumber Alfaria Trijaya Berbasis Barcode Scanner Android	Persediaan barang, Scanner Barcode, Sistem Informasi	barang, android, menu, data, tampilan, sistem, aplikasi, gambar, diagram, pada, barcode, sumber, persediaan, yang, alfaria, trijaya, dengan, masuk, untuk, diusulkan
Aplikasi Sistem Informasi Penyaluran Dana Bantuan Operasional Sekolah Berbasis Web Pada SD Negeri Cimone 4 Tangerang	Informasi, Asismen, dana bos	dana, sekolah, pendidikan, halaman, sistem, bantuan, data, informasi, operasional, diagram, tata, laporan, usaha, gambar, class, yang, case, kegiatan, program, anggaran
Penyederhanaan Bentuk Motif Patra Olanda Menggunakan Software Berbasis Vektor Untuk Desain Kertas Dinding	Patra Olanda, CorelDRAW, Desain Wallpaper, Ruangan	desain, gambar, patra, olanda, kertas, bali, dinding, hasil, yang, motif, ruangan, untuk, bentuk, coreldraw, wallpaper, vektor, bangunan, software, pola, bunga

Dari beberapa kali percobaan yang diuji secara manual dengan membandingkan kata kunci artikel dengan kata kunci yang didapatkan menggunakan *word frequency* didapatkan bahwa, kata kunci yang didapatkan dengan analisa frekuensi kemunculan kata yang sering muncul dapat digunakan untuk menghasilkan kata kunci yang mewakili isi artikel. Apabila dibandingkan kata kunci artikel dengan kata

kunci yang dihasilkan menggunakan frekuensi kata memiliki kecenderungan yang mendekati sama. Penelitian ini menghasilkan kata kunci dengan metode yang lebih sederhana dari yang pernah dibuat sebelumnya, hanya menggunakan analisis frekuensi kata.

4. Kesimpulan

Dari modul yang dibuat dalam bentuk perangkat lunak sederhana dan diuji dengan pengujian manual, didapat ditarik kesimpulan bahwa penelitian yang dibuat untuk menghasilkan kata kunci secara otomatis dengan menggunakan *word frequency* dapat dibuat dengan modul web sederhana dengan metode yang lebih sederhana dari penelitian sebelumnya. Hasil *word frequency* yang didapatkan dapat digunakan sebagai kata kunci sebuah artikel. Kata kunci yang dihasilkan menggunakan *word frequency* dapat dibandingkan dengan kata kunci artikel yang dituliskan dengan hasil yang mendekati sama. Untuk penelitian selanjutnya diperlukan pengujian tingkat akurasi kata kunci yang dihasilkan dan juga dapat dikembangkan dengan metode yang lain untuk meningkatkan tingkat akurasi hasil kata kunci yang didapatkan. Selain itu penelitian ini akan dilanjutkan dengan memilih tim penilai otomatis dengan menghitung tingkat kesamaan kata kunci artikel dengan kompetensi yang dimiliki oleh masing-masing tim penilai.

Daftar Pustaka

[1] A. A. Ma'arif, "Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah," *Teknik Informatika Universitas Dian Nuswantoro*, 2015.

[2] M. Alexandrov, A. Gelbukh, and P. Rosso, "An approach to clustering abstracts," in *International Conference on Application of Natural Language to Information Systems*, Berlin, 2005, pp. 275–285.

[3] S. Beliga, "Keyword extraction: a review of methods and approaches," *University of Rijeka*, pp. 1–9, 2014.

[4] M. Habibi and A. Popescu-Belis, "Keyword extraction and clustering for document recommendation in conversations," in *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 2015, pp. 746-759.

[5] U. Dewi, Imelda, "Analisis Text Mining, Algoritma TF/IDF (Term Frequency-Inversed Document Frequency) dan Algoritma Vector Space Model Pada Pengelolaan Materi Ajar," in *Konferensi Nasional Sistem & Informatika (KNS&I)*, 2014.

[6] L. Agusta, "Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia," in *Konferensi Nasional Sistem & Informatika (KNS&I)*, 2009, pp. 196-201.

[7] O. Nurdiana, Jumadi, and D. Nursantika, "Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al-Qur'an Dalam Bahasa Indonesia," *Jurnal Online Informatika*, vol. 1, no. 1, pp. 59–63, 2016.

[8] I. K. R. Y. Negara, "Implementasi Si Toni Sebagai Ujian Berbasis Komputer pada Bisma Informatika Indonesia," in *SEMNASTEKNOMEDIA ONLINE*, 2016, pp. 25–30.

[9] V. Vaswani, *How to do Everything with PHP & MySQL*, United States of America: McGraw-Hill, 2005.

[10] R. K. Roul, O. R. Devanand, and S. K. Sahay, "Web Document Clustering and Ranking Using TF-IDF Based Apriori Approach," in *IJCA Proceedings on ICACEA*, 2014, pp. 74–78.