Ekstraksi Judul dan Abstrak Artikel Ilmiah Berbasis Rule

Yosua Setyawan Soekamto, *Departemen Sistem Informasi, Universitas Ciputra Surabaya*

***Abstrak*—Perkembangan penelitian dan jumlah *research paper* yang dipublikasikan di berbagai Jurnal, menimbulkan kesulitan pada proses seleksi dan referensi oleh para peneliti dan pengelola jurnal. Dalam *research paper* bagian judul dan abstrak adalah ide utama dan ringkasan penelitian beserta metode yang digunakan dalam penelitian tersebut. Ekstraksi judul dan ringkasan *research paper* menjadi topik yang cukup banyak dibahas dengan berbagai metode. Umumnya ekstraksi terbatas dengan penggunaan bahasa dan gaya penulisan tiap-tiap jurnal. Dalam penelitian ini, dilakukan ekstraksi judul dan abstrak dengan menggunakan *association rule.* Penerapannya dilakukan dengan intuisi umum dalam penulisan *research paper*. Penelitian yang dilakukan menggunakan 2 dataset *layout research paper*, yaitu bentuk 1 kolom dan 2 kolom. Penelitian ini membantu pengelola jurnal dan peneliti sehingga memudahkan proses referensi secara otomatis dan proses seleksi untuk publikasi jurnal secara *online*. *Rule* diterapkan pada gaya penulisan *research paper* yang umum digunakan, sehingga dapat diterapkan pada berbagai jenis bahasa. Beberapa contoh rule yang digunakan adalah "Judul *paper* merupakan sebuah kalimat (frase) dengan menggunakan ukuran teks yang paling besar", "Judul *paper* ditulis pada awal halaman pertama", "Judul *paper* mayoritas ditulis dengan menggunakan cetak tebal (*bold*)", "Judul paper diikuti dengan nama penulis", "Judul *paper* yang muncul di halaman kedua dan selanjutnya sebagai *header* atau *footer* memiliki letak yang tidak lazim dibanding isi *paper* (atau berada di *margin* halaman)". Hasil penelitian memberikan 98% kesuksesan memperoleh Judul *paper*, 76% memperoleh Abstrak dalam Bahasa Indonesia dan 80% Abstact dalam Bahasa Inggris.**

***Kata Kunci*—Information Extraction, Association Rules, Title Extraction, Abstract Extraction, Text Metadata Extraction.**

I. Pendahuluan

Dalam meningkatkan kualitas kehidupan, banyak peneliti yang terus mengembangkan ilmu dengan melakukan penelitian di berbagai bidang. Untuk dapat disahkan dan diakui oleh masyarakat, maka peneliti menulis ringkasan hasil penelitiannya ke dalam dokumen yang dikenal sebagai *research paper*. *Research paper* ini selanjutnya dipublikasikan ke dalam jurnal, baik dalam skala lokal internal dalam negeri maupun skala internasional. Bagi seorang peneliti, tingkat kualitas *research paper* dinilai dari banyaknya referensi yang merujuk *papernya*. Sedangkan bagi pengelola jurnal, kualitas jurnal dinilai dari banyaknya *research paper* yang berkualitas dan juga mempengaruhi tingkat konsistensi publikasi jurnal tersebut.

Sistem ekstraksi umumnya menggunakan proses *machine learning* dan *natural language processing* (NLP), sedangkan pada penelitian ini digunakan *association rules* yang diterapkan pada gaya penulisan (*font style*) dan struktur penulisan *research paper* (*layout*).

II. Tinjauan Pustaka

Penelitian tentang ekstraksi pada research paper telah banyak dilakukan tetapi mayoritas menggunakan *natural language processing* (NLP) sebagai dasarnya [1]. Penggunaan NLP umumnya disertai juga dengan penggabungan *machine learning*, biasanya bertujuan untuk melatih sistem yang dibangun sehingga dapat beradaptasi dengan research paper yang baru. Sistem seperti ini membutuhkan proses yang relatif cukup lama, karena membutuhkan waktu untuk *training* dan beradaptasi, selain itu dibutuhkan juga cukup banyak dataset untuk melatihnya.

Penelitian yang lebih lanjut adalah dengan menggunakan metode *Neural Network* yaitu *Back-Propagation* [2]. Mulanya dilakukan metode *machine learning* pada teks di bagian abstrak dan judul dokumen. Dari teks-teks tersebut dilakukan pembobotan unigram, bigram dan trigram (n-kata), kemudian dicocokkan dengan *keyword* yang ditulis oleh penulis dokumen. Proses pencocokan dilakukan dengan dua cara yaitu *full-keyword* dan *partial-keyword*. *Full-keyword* berarti mencocokkan teks *unigram*, *bigram* dan *trigram* dengan teks *keyword* utuh dari penulis, termasuk kepanjangan-singkatan yang ditulis penulis, sedangkan partial-keyword adalah mencocokkan dengan n-karakter awal, misalnya *cryptogram*, *cryptography*, *cryptosystem* memiliki kesamana 5-karakter *crypt*. Selanjutnya dipilih *keyword* yang memiliki bobot yang tinggi.

Selanjutnya muncul ide baru penelitian yang menggunakan gabungan metode NLP, statistik, rule dan *back-propagation* [3]. Dalam penelitian tersebut, teks diambil setiap kata dan diberi label jenis kata (kata benda, kata sifat, kata kerja dan seterusnya), lalu dibentuk frase/kalimat kata benda (*noun-phrase* atau NP). Selanjutnya NP dicari pada judul dan abstrak, jika NP ada di judul maka diberi pembobotan j=1 dan sebaliknya jika ada pada abstrak maka a=1. Selanjutnya proses *rule-based* menyeleksi bobot NP sesuai kemunculan pada judul dan abstrak. Akhirnya hasil dari proses rule-based dimasukkan

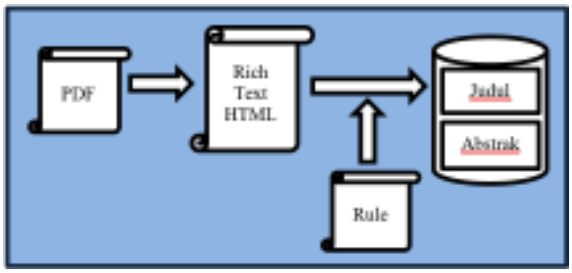
Maret 2020
Yosua Setyawan Soekamto, Departemen Sistem Informasi, Universitas Ciputra Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: yosua.soekamto@ciputra.ac.id)

pada proses back-propagation untuk menghitung bobot secara berulang hingga mencapai threshold yang diinginkan.

Fokus penelitian ini adalah merujuk dari karakteristik yang disimpulkan dari penelitian-penelitian rujukan yang disebutkan sebelumnya. Karakteristik pertama yaitu pencarian *keyword* difokuskan pada bagian judul dan abstrak dokumen. Judul dan abstrak merupakan ide pokok dan rangkuman dari isi dokumen [2],[4],[5]. Penggunaan *association rule* diharapkan dapat mempercepat proses ekstraksi dan menghemat *resource* spesifikasi komputer yang dipakai. *Rule* yang dibuat diadaptasi dengan struktur penulisan *research paper* seperti urutan-urutan isi *research paper*, atribut tiap-tiap teks dan bahasa yang digunakan [6],[7],[8].

III. METODE PENELITIAN

Metode penelitian menerapkan model rekayasa perangkat lunak (*software engineering*) secara umum. Awalnya dilakukan analisa pada dokumen-dokumen research paper, mengacu tiga karakteristik yaitu pada bagian judul dan abstrak dokumen, *threshold* untuk bobot dan pencocokan *keyword* pada seluruh dokumen; dan mencari karakteristik lain yang menjelaskan struktur penulisan research paper. Dari hasil analisa dibuat desain sistem dengan membentuk *rule-rule* sederhana dengan *threshold-threshold* yang bisa dimodifikasi sesuai kebutuhan. Kemudian dilakukan implementasi pembuatan *software* sesuai dengan *rule* dan *threshold* yang sudah didesain, dan melakukan uji coba pada beberapa dokumen *research paper*. Akhirnya dilakukan pemeliharaan sistem dengan melakukan pembenaran *threshold* dari hasil uji coba. Proses implementasi, uji coba dan pembenaran threshold ini berlangsung berulang-ulang, sampai ditemukan *threshold* yang sesuai. Proses ini memang memakan waktu cukup banyak pada saat pembuatan program dibandingkan metode *machine learning*, tetapi hanya berlaku pada awal pembuatan program ini saja. Jika sudah ditemukan *threshold* yang memadai, maka dipercaya sistem mampu mengolah dokumen-dokumen lebih cepat dibanding metode *machine learning*.



Gambar. 1. Desain Arsitektur Penelitian

Sistem pada mulanya mengubah teks dokumen PDF menjadi *rich-text* dokumen yang bisa berupa XML atau HTML. *Rich-text* dokumen yang dimaksud adalah dokumen dengan keterangan-keterangan format penulisan seperti ukuran teks dan hiasan teks (cetak tebal, cetak miring dan sebagainya). Fungsinya adalah mencari teks-teks tertentu yang menjadi judul dan abstrak. Sebagai gambaran umum, teks judul memiliki keunikan penulisan, seperti "ditulis dengan ukuran paling besar dan dicetak tebal" [8]. Keunikan-keunikan tersebut dituang dalam bentuk *association rule* yang juga menjadi fokus utama penelitian ini. Proses pengubahan format dari PDF menjadi XML atau HTML menggunakan PDFBox.

Contoh *association rules* yang digunakan antara lain:

1. "Judul *paper* ditulis di awal halaman pertama."
2. "Judul *paper* ditulis dengan ukuran teks paling besar."
3. "Judul *paper* bisa ditulis dengan menggunakan cetak tebal (*bold*)."
4. "Abstrak dimulai dengan kata kunci 'abstract' jika menggunakan bahasa Inggris."
5. "Abstrak yang menggunakan bahasa Indonesia dimulai dengan kata kunci 'abstrak' atau 'intisari'."
6. "*Paper* dapat ditulis dengan 1 kolom atau maksimal 2 kolom."
7. "Jika *paper* ditulis dengan 1 kolom, maka isi abstrak dipisahkan baris dengan kata kunci-nya."
8. "Jika *paper* ditulis dengan 2 kolom, maka isi abstrak menjadi satu kesatuan dengan kata kunci-nya."
9. "Jika *paper* ditulis dengan 2 kolom, maka bagian abstrak ditulis dengan menggunakan cetak tebal (bold)."
10. "Sub-bab dalam *paper* dipisahkan dengan jeda baris yang cukup jauh (*line spacing*)"
11. "Jarak *line spacing* dihitung dengan ukuran teks (*font size*) yang paling sering muncul, karena *font size* tersebut menjadi ukuran penulisan *paper*"
12. "*Line spacing* yang digunakan dalam *paper* adalah 1 spasi, kecuali pada judul sub-bab, keterangan gambar, dan keterangan tabel."

Contoh association rules yang ditulis dalam program adalah:

1. IF [node_html] is firstPage THEN add into [candidate_node_html]
2. IF [candidate_node_html].fontSize is maxFontSize THEN add into [candidate_arraylist_title]
3. IF [candidate_arraylist_title].fontStyle is boldStyle THEN add into [arraylist_title]
4. IF [candidate_node_html].text in [arraylist_abstracts] THEN add into [candidate_arraylist_abstract]
5. IF [document].layout is havingMoreColumns THEN [document] labeled as twoColumns
6. IF [document].layout isNot havingMoreColumns THEN [document] labeled as oneColumn
7. IF [document] is oneColumn THEN findNextParagraph on [candidate_arraylist_abstract], add into [arraylist_abstract]
8. IF [document] is twoColumns THEN add [candidate_arraylist_abstract] into [arraylist_abstract]
9. IF lineDistance([node_html].top, [next_node_html].top) is lowerThan findMostFontSize([node_html]) THEN groupLine([node_html], [next_node_html])
10. IF wordDistance([node.html].left, [next_node_html].left) is lowerThan findMostFontSize[node_html]) THEN groupWord([node_html], [next_node_html])

(A) (B)

Gambar 2. Bentuk Paper 1 Kolom (A) dan Paper 2 Kolom (B)

IV. HASIL PENELITIAN

Pada penelitian ini digunakan dua jenis dataset yang dibedakan berdasarkan format penulisan research paper (1 kolom atau 2 kolom). Total *research paper* yang digunakan adalah 50 buah file PDF. PDF diubah menjadi HTML dengan menggunakan *open source tool* PDFBox yang berbasis bahasa pemrograman Java. Contoh dataset dapat dilihat pada gambar 2, (A) merupakan contoh *paper* dengan 1 kolom dan (B) merupakan contoh *paper* dengan 2 kolom.

Untuk menilai kemampuan sistem dari penelitian ini, dilakukan pengecekan secara manual terhadap 50 *research paper* pada dataset. Pencocokan manual ini bertujuan untuk menghitung nilai *accuracy* atau *error rate* dari sistem ekstraksi yang dibuat.

TABEL I
HASIL UJI COBA

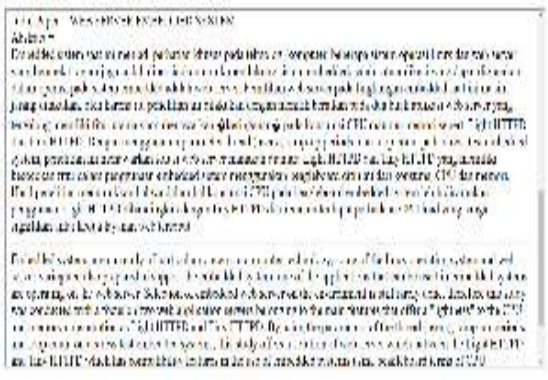
Keterangan	Jumlah	**Accuracy (%)**
Judul Paper		
Benar Ekstrak	49	**98%**
Salah Ekstrak	1	**2%**
Total	50	**100%**
Abstrak (Bahasa Indonesia)		
Benar Ekstrak	38	**76%**
Salah Ekstrak	12	**24%**
Total	50	**100%**
Abstract (Bahasa Inggris)		
Benar Ekstrak	40	**80%**
Salah Ekstrak	10	**20%**
Total	50	**100%**

Dataset yang digunakan adalah *research paper* yang menggunakan bahasa Indonesia, tetapi memiliki dua bagian abstrak yang ditulis dalam bahasa Indonesia dan bahasa Inggris. Pemilihan dataset berdasarkan tujuan penelitian, yaitu membuat sistem ekstraksi yang tidak sepenuhnya bergantung pada ragam bahasa.

Dari hasil uji coba, ditemukan bahwa ekstraksi judul *research paper* memiliki *accuracy rate* yang sangat tinggi, yaitu 98%, sedangkan perolehan ekstraksi abstrak berbahasa Indonesia sebesar 76% dan abstrak berbahasa Inggris sebesar 80%.

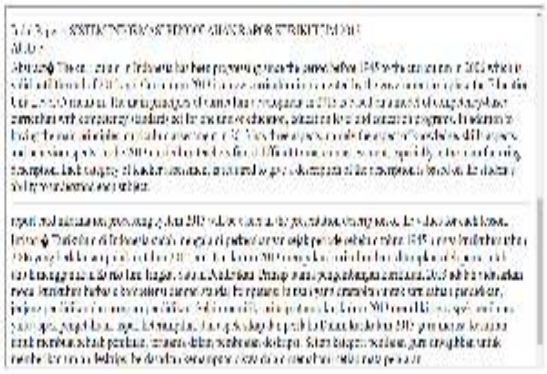
Sistem yang dibangun pada penelitian ini tergolong sukses melakukan ekstraksi judul dan abstrak dari research paper. Sistem ini mampu melakukan ekstraksi judul dan abstrak dari dua bentuk *paper* yang sangat umum digunakan (1 kolom dan 2 kolom). Contoh *paper* yang digunakan dapat dilihat pada Gambar 2, sedangkan hasil dapat dilihat pada Gambar 3.

Dari hasil pengamatan uji coba bagian abstrak memiliki keterbatasan jika menggunakan bentuk 2 kolom. Penyusunan teks dan pencarian kata kunci lebih sulit jika menggunakan bentuk 2 kolom. Selain itu pada percobaan pada dataset lain, ditemukan abstrak yang terpisah kolom, dan sistem ekstraksi penelitian ini tidak dapat menyatukan isi abstrak tersebut.



Gambar 3. Output dari gambar 2(A).

Contoh output yang didapatkan dari sistem ekstraksi penelitian ini terkait dengan contoh pada gambar 2 dapat dilihat pada gambar 3 dan gambar 4. Gambar 3 merupakan output dari research paper pada gambar 2(A). Judul paper yang didapatkan adalah "WEB SERVER EMBEDDED SYSTEM" dan abstrak dalam bahasa Indonesia ditampilkan pada paragraf atas dan abstrak dalam bahasa Inggris ditampilkan pada paragraf bawah. Dalam gambar 3 juga terlihat ada beberapa karakter yang tidak dapat diolah *encoding*nya.



Gambar 4. Output dari gambar 2(B).

Output dari contoh gambar 2(B) ditampilkan pada gambar 4. Hasil judul *paper* yang diperoleh adalah "SISTEM INFORMASI PENGOLAHAN RAPOR KURIKULUM 2013" dan abstrak yang diperolehpun ditampilkan dalam 2 paragraf. Paragraf yang atas berisi abstrak dalam bahasa Inggris dan paragraf yang bawah berisi abstrak dalam bahasa Indonesia. Pada contoh dataset ini, abstrak bahasa Indonesia diawali dengan kata kunci "Intisari" dan pada ekstraksi ini juga ditemui karakter yang encodingnya tidak dapat diolah. Untuk karakter yang tidak dapat diolah *encoding*nya tampil sebagai simbol "**"**.

Kegagalan ekstraksi pada abstrak disebabkan kesulitan pembatasan paragraf atau segmen bab pada *paper*. *Rule set* yang digunakan untuk membatasi antar bab/bagian *paper* terbukti gagal. Kegagalan ini mengakibatkan sistem tidak dapat menyusun ulang abstrak sehingga dikategorikan sebagai uji coba gagal. *Rule set* yang digunakan adalah mengukur jarak teks dengan baris atau bagian selanjutnya. *Threshold* untuk pengukuran jarak ini harus diteliti lebih lanjut sehingga dapat memberikan akurasi yang lebih baik.

V. KESIMPULAN

Kesimpulan yang diperoleh dari penelitian ini antara lain:

1. Sistem yang dibangun pada penelitian ini mampu melakukan ekstraksi judul dan abstrak dengan tingkat accuracy yang tinggi, yaitu 98% untuk memperoleh Judul, 76% memperoleh abstrak berbahasa Indonesia dan 80% memperoleh abstrak berbahasa Inggris.
2. Pendekatan menggunakan *association rules* pada ekstraksi judul dan abstrak terbukti relatif lebih cepat dan memiliki tingkat *accuracy* yang relatif cukup tinggi.
3. Penggunaan PDFBox harus diikuti dengan pre-proses penyusunan teks kembali, karena cara kerja PDFBox menggunakan segmentasi blok pada teks PDF.
4. Kata kunci untuk mensegmentasi bagian-bagian paper sebaiknya disimpan dalam bentuk *text-file* sehingga dapat dimodifikasi sesuai dengan penggunaan bahasa pada dataset.
5. *Encoding* yang digunakan pada PDFBox perlu diperhatikan karena terdapat beberapa karakter khusus yang tidak dapat diolah dengan basis teks.
6. Pada penelitian ini penggunaan association rules tidak menggunakan proses *scoring* maupun *ranking*, tetapi hasil bisa saja meningkat jika diberikan scoring pada ekstraksi bagian-bagian isi paper.
7. *Threshold* jarak antar baris/bagian *paper* perlu dilakukan perhitungan lebih lanjut, karena akan mempengaruhi akurasi ekstraksi abstrak.

Penelitian ini merupakan bagian awal untuk penelitian keyword extraction dari *research paper*, sehingga penelitian ini dapat dilanjutkan dengan menggunakan metode NLP dan machine learning untuk mencari kata dasar sebagai bagian keyword extraction. Saran untuk pengembangan selanjutnya adalah meningkatkan *accuracy* pada bagian abstrak, sehingga dapat mendukung proses keyword extraction.

Keyword extraction yang disinggung pada tinjauan pustaka umumnya melakukan penelusuran pada seluruh teks *research paper*. Penulis mengusulkan melakukan penelusuran pada bagian judul *paper*, abstrak dan metode yang digunakan pada *research paper*. Ketiga bagian tersebut merupakan fokus utama dari research paper, oleh karena itu proses keyword extraction diusulkan dicari pada tiga bagian tersebut, tetapi perlu diwaspadai sistem ekstraksi terjebak pada bagian tinjauan pustaka atau pendahuluan yang umumnya juga memaparkan sebagian metode-metode tetapi bukan metode penelitian utamanya.

ACKNOWLEDGEMENT

Penulis mengucapkan terima kasih pada Lembaga Penelitian dan Pengabdian Masyarakat Universitas Ciputra Surabaya yang telah membantu mendanai penelitian ini. Penulis juga mengucapkan terima kasih kepada Fakultas Teknologi Informasi yang telah membantu penelitian ini sebagai anggota pakar dalam pengumpulan data.

DAFTAR PUSTAKA

[1] Frank, Eibe., Witten, Ian H., Paynter, Gordon W., Gutwin, Carl., Nevill-Manning ,Craig G., "Domain Specific Keyphrase Extraction", Proceedings 16th International Joint Conference on Artificial Intelligence. 1999.

[2] Bhowmik, Rekha., "Keyword Extraction from Abstracts and Titles", Proceedings of the IEEE Southeastcon. 2008.

[3] Kavila, Selvani Deepthi, Rajesh, B., Vyshnavi, N., Sushma, K. Moni., "Automatic Key Term Extraction from Research Article using Hybrid Approach", International Journal of Computer Application, Volume 166 No. 6, May. 2017.

[4] Kaur, Jasmeen., Gupta, Vishal., "Effective Approaches for Extraction of Keywords", International Journal of Computer Science, Volume 7, Nov. 2010.

[5] Rose, Stuart., Engel, Dave., Cramer, Nick., Cowley, Wendy., "Automatic Keyword Extraction from Individual Documents", Text Mining: Applications and Theory, 2010.

[6] Guo, Zhixin., Jin, Hai., "A Rule-Based Framework of Metadata Extraction from Scientific Papers", 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, 2011.

[7] Soderland, S., "Learning Information Extraction Rules for Semi-Structured and Free Text", Kluwer Academic Publishers, 1999.
[8] Beel, Joran., Gipp, Bela., Shaker, Ammar., Friedrich, Nick., "SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size)", Proceedings of the 14th European Conference on Digital Libraries, Volume 6273, Sept. 2010.
[9] Hasan, H. M. Mahedi., Sanyal, Falguni., Chaki, Dipankar., Ali, Md. Haider., "An Empirical Study of Important Keyword Extraction Techniques from Documents", International Conference on Intelligent System and Information Management, Oct. 2017.
[10] Matsuo, Y., Ishizuka, M., "Keyword Extraction from a Single Document using Word Co-Occurrence Statistical Information", International Journal on Artificial Intelligence Tools, Dec. 2003.
[11] Zhang, Chengzhi., Wang, Huilin., Liu Yao., Wu, Dan., Liao, Yi., Wang, Bo., "Automatic Keyword Extraction from Document Using Conditional Random Fields", Journal of Computational Information Systems, 2008.
[12] Witten, Ian H., Paynter, Gordon W., Frank, Eibe., Gutwin, Carl., Nevill-Manning ,Craig G., "KEA: Practical Automatic Keyphrase Extraction", in Proceedings of the 4th ACM Conference on Digital Libraries, 1998.

Yosua Setyawan Soekamto lahir di Surabaya, Jawa Timur, pada tahun 1990. Dia menyelesaikan studi S1 Teknik Informatika Sekolah Tinggi Teknik Surabaya pada tahun 2012 dan studi master Teknologi Informasi Sekolah Tinggi Teknik Surabaya pada tahun 2017. Minat penelitiannya adalah bidang *software engineering*, *software design*, *web development*, *information systems* dan sedang dalam pembelajaran *mobile device programming*.