Pengaruh *Query Expansion* Terhadap Pendeteksian Kemiripan Teks Menggunakan *Cosine Similarity*

Diajukan untuk Menyusun Tugas Akhir
di Jurusan Teknik Informatika Fakultas Ilmu Komputer Unsri

Oleh :

Pipit Kurnia Sari
NIM : 09021181520025

JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2019

LEMBAR PENGESAHAN TUGAS AKHIR

PENGARUH *QUERY EXPANSION* TERHADAP PENDETEKSIAN KEMIRIPAN TEKS MENGGUNAKAN *COSINE SIMILARITY*

Oleh :

PIPIT KURNIA SARI
NIM : 09021181520025

Indralaya, 19 September 2019

Pembimbing I,

Novi Yusliani, M.T
NIP. 198211082012122001

Pembimbing II,

Yunita, M.Cs
NIP. 198306062015042002

Mengetahui,
Ketua Jurusan Teknik Informatika.

Rifkie Primartha, M.T
NIP. 197706012009121004

TANDA LULUS UJIAN SIDANG TUGAS AKHIR

Pada hari tanggal 19 September 2019 telah dilaksanakan ujian sidang tugas akhir oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Pipit Kurnia Sari
NIM : 09021181520025
Judul : Pengaruh *Query Expansion* Terhadap Pendeteksian Kemiripian Teks Menggunakan *Cosine Similarity*

1. Pembimbing I

   Novi Yusliani, M.T
   NIP. 198211082012122001 .....

2. Pembimbing II

   Yunita, M.Cs
   NIP. 198306062015042002 .....

3. Penguji I

   M. Fachrurrozi, M.T
   NIP. 198005222008121002 .....

4. Penguji II

   Kanda Januar Miraswan, M.T
   NIP. .....

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T
NIP. 197706012009121004

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Pipit Kurnia Sari
NIM : 09021181520025
Program Studi : Teknik Informatika
Judul Skripsi : Pengaruh *Query Expansion* Terhadap Pendeteksian Kemiripan Teks Menggunakan *Cosine Similarity*
Hasil Pengecekan Software *iThenticate/Turnitin* : 16%

Menyatakan bahwa Laporan Projek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan projek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Indralaya, September 2019

Pipit Kurnia Sari
NIM. 09021181520025

Motto :

- *Menggapai ridha Allah dan orang tua*
- *Percaya Rencana Allah Yang Terbaik*

Kupersembahkan karya tulis ini kepada :

- *Orang tuaku tercinta*
- *Kakak, Ayuk dan Adikku tersayang*
- *Keluarga besarku*
- *Sahabat dan teman seperjuanganku*
- *Fakultas Ilmu Komputer*
- *Universitas Sriwijaya*

THE EFFECT OF QUERY EXPANSION AGAINST DETEVERATION OF TEXT-RELATED USING COSINE SIMILARITY

By :
Pipit Kurnia Sari
09021181520025

ABSTRACT

Cosine Similarity is a method of calculating the similarity of text that depends on the same word as the word being tested. If the word in the test text is not the same as the word in the source text, then the word does not match the word in the word list and the word cannot be counted. This research examines the effect of query expansion using a thesaurus, which is one algorithm to improve the effectiveness of a word list match. Cosine similarity algorithm with query expansion or without query expansion each tested with 7 source documents and 21 comparison documents. Based on cosine similarity evaluation results with query expansion can improve the detection of text similarity compared to the cosine similarity algorithm without query expansion, which is a percentage value of 46.90%, on data without query expansion and 43.11% for window size 2, 42.90 % for window size 3, 42.59% for window size 4. Although it can increase overall computing time, however, the term obtained from forming query expansion makes the text similarity better.

Keywords: Expansion Query, Thesaurus, Cosine Similarity, Window Size.

Indralaya, 19 September 2019

Pembimbing I,

Novi Yusliani, M.T
NIP. 198211082012122001

Pembimbing II,

Yunita, M.Cs
NIP. 198306062015042002

Mengetahui,
Ketua Jurusan Teknik Informatika,

Rifkie Primartha, M.T
NIP.197706012009121004

PENGARUH *QUERY EXPANSION* TERHADAP PENDETEKSIAN KEMIRIPAN TEKS MENGGUNAKAN *COSINE SIMILARITY*

Oleh:
Pipit Kurnia Sari
09021181520025

ABSTRAK

Cosine Similarity merupakan salah satu metode untuk menghitung kemiripan teks yang bergantung pada kata yang sama dengan kata yang diujikan. Apabila kata yang ada pada teks uji tidak sama dengan kata pada teks sumber, maka kata tersebut tidak cocok dengan kata yang ada dalam daftar kata dan kata tersebut tidak dapat dihitung. Pada penelitian ini menguji pengaruh *query expansion* menggunakan *thesaurus* yang merupakan salah satu algoritma untuk meningkatkan keefektifan dari kecocokan daftar kata. Algoritma *cosine similarity* dengan *query expansion* maupun tanpa *query expansion* masing-masing diuji dengan 7 dokumen sumber dan 21 dokumen pembanding. Berdasarkan hasil evaluasi *cosine similarity* dengan *query expansion* mampu meningkatkan pendeteksian kemiripan teks dibandingkan dengan algoritma *cosine similarity* tanpa *query expansion*, yaitu dengan nilai persentase sebesar 46,90%, pada data tanpa *query expansion* dan 43,11% untuk *window size* 2, 42,90% untuk *window size* 3, 42,59% untuk *window size* 4. Meskipun dapat menambah waktu komputasi secara keseluruhan namun, term yang diperoleh dari pembentukan *query expansion* membuat kemiripan teks semakin bagus.

Kata Kunci: *Query Expansion, Thesaurus, Cosine Similarity, Window Size*.

Indralaya, 19 September 2019

Pembimbing I,

Novi Yusliani, M.T
NIP. 198211082012122001

Pembimbing II,

Yunita, M.Cs
NIP. 198306062015042002

Mengetahui,
Ketua Jurusan Teknik Informatika,

Rifkie Primartha, M.T
NIP.197706012009121004

KATA PENGANTAR

Bismillahirrahmanir rahiim

Puji syukur kepada Allah atas berkat dan rahmat-Nya yang telah diberikan kepada Penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan baik. Tugas akhir ini disusun untuk memenuhi salah satu syarat guna menyelesaikan pendidikan program Strata-1 pada Fakultas Ilmu Komputer Program Studi Teknik Informatika di Universitas Sriwijaya.

Dalam menyelesaikan Tugas Akhir ini banyak pihak yang telah memberikan bantuan dan dukungan baik secara langsung maupun secara tidak langsung. Penulis ingin menyampaikan rasa terima kasih kepada:

1. Orang tuaku, Sarkowi Hasan dan Sujirah, kakakku, Ahyarrudin, ayukku, Desi Kusuma Dewi, dan adikku, Febri Rendi Irawa serta seluruh keluarga besarku yang selalu mendokan serta memberikan dukungan baik moril maupun materil.
2. Bapak Jaidan Jauhari, M.T selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya, Bapak Rifkie Primartha, M.T selaku Ketua Jurusan Teknik Informatika dan Ibu Hardini Novianti, M.T selaku Sekretaris Jurusan Teknik Informatika.
3. Ibu Novi Yusliani, M.T selaku dosen pembimbing I dan Ibu Yunita, M.Cs selaku pembimbing II, yang telah membimbing, mengarahkan dan memberikan motivasi penulis dalam proses perkuliahan dan pengerjaan Tugas Akhir.
4. Ibu Novi Yusliani, M.T selaku dosen pembimbing akademik, yang telah membimbing, mengarahkan dan memberikan motivasi penulis dalam proses perkuliahan dan pengerjaan Tugas Akhir.
5. Bapak M. Fachrurrozi, M.T selaku dosen penguji I, dan Bapak Kanda Januar Miraswan, M.T selaku dosen penguji II yang telah memberikan masukan dan dorongan dalam proses pengerjaan Tugas Akhir.
6. Seluruh dosen Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Kak Ricy serta seluruh staf tata usaha yang telah membantu dalam kelancaran proses administrasi dan akademik selama masa perkuliahan.
8. Achi, Lifya, Sari, Yulinda, Ayu, Puli serta seluruh teman-teman jurusan Teknik Informatika yang telah saling berbagi selama masa perkuliahan ini.
9. DPM KM Fasilkom, Wifi yang telah memberikan ruang bagi Penulis untuk berprestasi dan berkarya.

Penulis menyadari dalam penyusunan Tugas Akhir ini masih terdapat banyak kekurangan disebabkan keterbatasan pengetahuan dan pengalaman, oleh karena itu kritik dan saran yang membangun sangat diharapkan untuk kemajuan penelitian selanjutnya. Akhir kata semoga Tugas Akhir ini dapat berguna dan bermanfaat bagi kita semua.

Indralaya, September 2019

Pipit Kurnia Sari

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

Halaman

BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada bab ini membahas latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian serta batasan masalah. Bab ini akan memberikan penjelasan umum mengenai keseluruhan penelitian.

Pendahuluan dimulai dengan penjelasan mengenai tantangan dan tujuan proses menemukan pengetahuan baru pada deteksi kemiripan teks. Serta penelitian yang berkaitan dengan menerapkan *Query Expansion* yang menjadi latar belakang dari penelitian ini.

1.2 Latar belakang

Dalam perkembangan dunia teknologi informasi plagiarisme secara umum mengacu pada penyalinan informasi atau menggandakan karya seseorang yang tidak diketahui sumbernya, seperti dokumen dan program(Muhammad *et al.*, 2017). Serta penggunaan ulang materi miliknya sendiri (dikenal sebagai *self-plagiarism),* dan yang dihasilkan oleh orang lain(Muhammad *et al.*, 2017). Terutama dalam pendidikan tinggi, plagiarisme diakui sebagai masalah yang signifikan dan telah dilaporkan semakin meningkat (Park, 2003). Misalnya, (Citron and Ginsparg, 2015) menganalisis penggunaan kembali teks dalam korpus ilmiah ArXiv.org. Akibatnya, plagiarisme dan pendeteksiannya baru-baru ini menerima perhatian yang signifikan (Boisvert and Irwin, 2006). Sehingga perlu adanya tindakan pendeteksian

kemiripan teks dari karya-karya tulis supaya nilai keaslian dari teks tersebut dapat diketahui (Ryansyah and Andayani, 2017).

Berbagai faktor dapat menandakan plagiarisme, seperti referensi yang salah dan kesamaan yang sama dengan materi yang ada. Secara umum, pendekatan untuk mendeteksi plagiarisme (baik manual atau otomatis) dapat dikategorikan ke dalam dua masalah utama. Deteksi plagiarisme intrinsik berhubungan dengan mengindentifikasi inkonsistensi gaya dalam teks yang menimbulkan pertanyaan tentang kepenulisannya. Detekti plagiarisme ekstrinsik berkaitan dengan mengindentifikasi kemungkinan sumber dari dokumen yang mencurigakan (Stein, Eissen and Potthast, 2007). Maka pada penelitian ini menggunakan deteksi plagiarisme ekstrinsik.

Penelitian mengenai pendeteksian kemiripan teks bahasa Indonesia telah banyak diteliti sebelumnya. Salah satu metode yang digunakan adalah metode *Cosine Similarity*. Pada metode tersebut, didapatkan persentase kemiripan suatu teks berdasarkan perhitungan jumlah kemunculan kata pada teks pembanding (Imbar et al., n.d. 2014). Dalam penelitian lain yang dilakukan oleh (Firdaus, Ernawati and Vatresia, 2014) menjelaskan setiap kata harus diubah menjadi kata dasar terlebih dahulu pada tahap *preprocessing* sebelum melakukan perhitungan kemiripann teks. Hasil pengujian dari percobaan tersebut dihitung kemiripan teksnya menggunakan *Cosine Similarity* didapatkan persentase tingkat akurasi sebesar 87,83%, sedangkan untuk hasil pengujian setelah dilakukan tahap *Praprocessing* persentase tingkat akurasi menjadi 93,81% (Firdaus, Ernawati and Vatresia, 2014).

Salah satu faktor kendala menghitung kesamaan teks adalah ketika kata pada teks uji berbeda dengan kata pada teks yang aslinya. Kata tersebut tidak dapat dihitung. Kegagalan ini disebut ketidakcocokan daftar kata atau *vocabulary mismatch* (Carpineto and Romano, 2012). Untuk menutupi kendala ini, maka diperlukannya suatu metode untuk meningkatkan keefektifan dari ketidakcocokan daftar kata atau ketidakkonsistenan pada pengindeksan dokumen yaitu dengan menerapkan teknik *Query Expansion* menggunakan *Thesaurus*. *Thesaurus* dapat memecahkan masalah ketidakkonsistenan pada pengindeksan dokumen, dan juga dapat digunakan dengan pencarian dalam memformulasi ulang strategi pencarian yang tepat jika diperlukan (Cholifah, Purwanto and Bramanto, 2011). *Thesaurus* akan menyediakan daftar kata yang tepat dan terkontrol yang menunjukkan keterkaitan istilah dan dapat digunakan sebagai alat untuk memperluas *query*, juga dalam mengkoordinasikan pengindeksan maupun pencarian dokumen (Khafajeh, Refai and Yousef, 2013).

Rasyidi, Romadhony and Wibowo, (2013) dalam penelitiannya tentang sistem temu kembali informasi *hadits* menerapkan *Query Expansion* menggunakan *Thesaurus*. Pengujian tingkat akurasi dari hasil penelitian tersebut menggunakan MAP *(Mean Average Precision)* dan *Recall*. Dalam penelitian ini adanya peningkatan performansi sistem temu kembali informasi sebelum dan sesudah menggunakan *Thesaurus* dalam proses *Query Expansion* dengan peningkatan MAP sebesar 34% dan *Recall* sebesar 43%. Penelitian yang dilakukan oleh Muhammad et al. (2017) dalam penelitiannya menerapkan *Query Expansion* menggunakan UMLS Metathesaurus dan *MEDLINE*, pada sistem pendeteksian kemiripan teks.

Pendekatan berbasis IR yang diusulkan di sini yaitu *Cosine Similarity* mencapai hasil yang lebih tinggi daripada pendekatan *Kullback-Leibler Distance*. Penarikan tertinggi yang dicapai dengan metode *Kullback-Leibler Distance* adalah 0,8596 untuk 20 dokumen kandidat teratas, meskipun diharapkan kinerja akan turun ketika seluruh basis data MEDLINE digunakan. Pendekatan yang diusulkan (dengan *query* expansion dan WSD) mencapai penarikan sebesar 0,9077 untuk 1 dokumen, yang masih lebih tinggi dari penarikan maksimum yang diperoleh menggunakan metode Kullback-Leibler Distance.

Dari penelitian-penelitian yang telah dijabarkan diatas, sistem pendeteksian kemiripan teks menggunakan teknik *Query Expansion* dapat memberikan hasil tingkat akurasi yang lebih baik dibandingkan dengan sistem pendeteksian kemiripan teks tanpa menggunakan *Query Expansion*. Sehingga sistem pendeteksian kemiripan teks bahasa Indonesia dengan menerapkan *Query Expansion* menggunakan *Thesaurus* diharapkan mampu meningkatkan kinerja sistem pendeteksian kemiripan teks dalam menyajikan informasi lebih cepat dan akurat.

1.3 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah jelaskan, rumusan masalah pada penelitian ini adalah bagaimana pengaruh *Query Expansion* menggunakan *Thesaurus* terhadap tingkat akurasi sebuah sistem pendeteksian kemiripan teks bahasa Indonesia menggunakan *Cosine Similarity* ?

Untuk menjawab rumusan masalah tersebut, dibawah ini diuraikan beberapa *research question* sebagai berikut :

1. Bagaimana mekanisme *Cosine Similarity* dalam sistem pendeteksian kemiripan teks bahasa Indonesia?
2. Bagaimana mekanisme *Query Expansion* menggunakan *Thesaurus* dalam sistem pendeteksian kemiripan teks bahasa Indonesia menggunakan *Cosine Similarity*?
3. Bagaimana hasil persentase kemiripan teks dalam sistem pendeteksian kemiripan teks bahasa Indonesia dengan menerapkan *Cosine Similarity* dan *Query Expansion?*
4. Bagaimana hasil persentase kemiripan teks dalam sistem pendeteksian kemiripan teks bahasa Indonesia menggunakan *Cosine Similarity* tanpa *Query Expansion?*
5. Bagaimana hasil perbandingan *Cosine Similarity* dengan *Query Expansionon* dan *Cosine Similarity* tanpa *Query Expansionon* pada sistem pendeteksian kemiripan teks bahasa indonesia berdasarkan dari hasil persentase?

1.4 Tujuan penelitian

Tujuan dari penelitian ini adalah sebagai berikut :

1. Mengetahui mekanisme *Cosine Similarity* dalam sistem pendeteksian kemiripan teks bahasa Indonesia.

2. Mengetahui mekanisme *Query Expansion* menggunakan *Thesaurus* dalam sistem pendeteksian kemiripan teks bahasa Indonesia menggunakan *Cosine Similarity*.
3. Dapat mengetahui pengaruh *Query Expansion* yang diterapkan pada sistem pendeteksian kemiripan teks menggunakan *Cosine Similarity* dengan melihat hasil persentase kemiripan teks.
4. Dapat mengetahui pengaruh *Query Expansion* jika tidak diterapkan pada sistem pendeteksian kemiripan teks menggunakan *Cosine Similarity* dengan melihat hasil persentase kemiripan teks.
5. Menganalisa hasil perbandingan berdasarkan hasil persentase pada metode *Cosine Similarity* dengan *Query Expansionon* dan *Cosine Similarity* tanpa *Query Expansionon* pada sistem pendeteksian kemiripan teks bahasa Indonesia.

1.5 Manfaat penelitian

Adapun manfaat yang diperoleh dalam penelitian ini adalah sebagai berikut :

1. Memahami *Query Expansion* menggunakan *Thesaurus* sebagai metode pendeteksian kemiripan teks bahasa Indonesia.
2. Hasil penelitian dapat dijadikan sebagai rujukan bagi peneliti lain dalam mengembangkan sistem pendeteksian kemiripan teks bahasa Indonesia.

1.6 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut :

1. Data teks yang digunakan berupa teks (huruf) berbahasa Indonesia, Data teks tidak memperhitungkan *equation* (rumus), tabel, simbol dan gambar.
2. data teks tidak menggunakan singkatan, sudah sesuai dengan ejaan yang disempurnakan (EYD) dan kata-katanya sudah baku.
3. Metode pembobotan kata yang digunakan adalah TF-IDF.
4. Metode pendeteksian kemiripan teks yang digunakan adalah *Query Expansion* dan *Cosine Similarity*.

1.7 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut :

BAB I. PENDAHULUAN

Bab I ini menguraikan latar belakang, perumusan masalah, tujuan dan manfaat penelitian serta batasan masalah, dan sistematika penulisan penelitian.

BAB II. KAJIAN LITERATUR

Pada bab II ini akan membahas landasan teori yang digunakan dalam penelitian ini, seperti analisis *preprocessing*, *query expansion*, *thesaurus*, *term co-occurrence*, analisis metode *cosine similarity*, *term frequency/inverse document frequency*. Selain itu di bab ini akan dibahas penelitian-penelitian lain yang relevan dengan peneitian ini.

BAB III. METODOLOGI PENELITIAN

Pada bab III ini akan membahas mengenai tahapan yang akan dilaksanakan pada penelitian ini. Pada tahapan penelitian ini akan dibahas dengan rinci dengan mengacu pada suatu kerangka kerja. Serta pada bagian akhir bab ini akan dijabarkan perancangan manajemen proyek perangkat lunak untuk pelaksanaan penelitian ini.

1.8 Kesimpulan

Pada penelitian ini berfokus pada deteksi plagiarisme ekstrinsik, dan untuk melihat ada tidaknya pengaruh dari menggunakan *Query Expansion* menggunakan *Thesaurus* terhadap pendeteksian kemiripan teks bahasa Indonesia. Metode yang digunakan untuk mendeteksi kemiripan teks bahasa Indonesia adalah *Query Expansion (Thesaurus)* dan *Cosine Similarity.* Kemudian pengujian akan dilakukan dengan melihat persentase keakurasian hasil kemiripan teks dari sistem pendeteksian kemiripan teks bahasa Indonesia dengan *Query Expansion* menggunakan *Thesaurus.* Kemudian hasil pengujian akan dibandingkan dengan hasil dari yang tidak menerapkan *Query Expansion* menggunakan *Thesaurus.*

DAFTAR PUSTAKA

Boisvert, R. F. and Irwin, M. J. (2006) 'Plagiarism on the rise', *Communications of the ACM*, 49(6), p. 23. doi: 10.1145/1132469.1132487.

Carpineto, C. and Romano, G. (2012) 'A Survey of Automatic Query Expansion in Information Retrieval', *ACM Computing Surveys*, 44(1), pp. 1–50. doi: 10.1145/2071389.2071390.

Cholifah, Purwanto, Y. and Bramanto, A. (2011) 'Aplikasi Information Retrieval untuk pembentukan Tesaurus Berbahasa Indonesia secara otomatis'. Surabaya, pp. 41–48.

Citron, D. T. and Ginsparg, P. (2015) 'Patterns of text reuse in a scientific corpus', *Proceedings of the National Academy of Sciences*, 112(1), pp. 25–30. doi: 10.1073/pnas.1415135111.

Firdaus, A., Ernawati and Vatresia, A. (2014) 'Aplikasi Pendeteksi Kemiripan pada Dokumen Teks Menggunakan Algoritma Nazief & Andriani Dan Metode Cosine Similirity', *Jurnal Teknologi Informasi*, 10(April), pp. 96–109.

Gusmita, R. H. *et al.* (2014) 'A rule-based question answering system on relevant documents of Indonesian Quran Translation', *2014 International Conference on Cyber and IT Service Management, CITSM 2014*, pp. 104–107. doi: 10.1109/CITSM.2014.7042185.

Imbar, R. V. *et al.* (no date) 'Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks', pp. 31–42.

Khafajeh, H., Refai, M. and Yousef, N. (2013) ‘Building Arabic Automatic Thesaurus Using Co-occurence Technique’, *Proceedings of International Conference on Communication, Media, Technology and Design*, (April 2013), pp. 28–32.

Lei, K., Tang, H. and Zeng, Y. (2018) ‘Keywords Extraction via Multi-relational Network Construction Keywords Extraction via Multi-relational Network’, (January 2013). doi: 10.1007/978-3-319-00951-3.

Muhammad, R. *et al.* (2017) ‘This is a repository copy of An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE. An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE’, *IEEE/ACM Transactions on Computational Biology and Bioinformatics JOURNAL OF COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, 14(4), pp. 796–804. doi: 10.1109/TCBB.2016.2542803.

Mutiara, A. B. and Agustina, S. (2008) ‘Anti Plagiarism Application with Algorithm Karp-Rabin at Thesis in Gunadarma University’, *arXiv preprint arXiv:0811.4349*, p. 9. Available at: http://arxiv.org/abs/0811.4349.

Park, C. (2003) ‘Assessment & evaluation in higher education in other (people’ s) words : plagiarism by university students--literature and’, *Assessment & Evaluation in Higher Education*, 28(5), pp. 241–288. doi: 10.1080/0260293032000120352.

Pressman, R. S. and Maxim, B. R. (2005) ‘Software Engineering’, p. 976.

Purwarianti, A. and Yusliani, N. (2012) ‘Sistem Question Answering Bahasa Indonesia untuk Pertanyaan Non-Factoid’, *Jurnal Ilmu Komputer dan Informasi*, 4(1), p. 10. doi: 10.21609/jiki.v4i1.151.

Rafles, A. (2013) ‘Plagiarisme Dokumen Dengan Pendekatan K-Gram Berbasis Frasa K-Gram Berbasis Frasa’. doi: 10.1186/1478-4491-13-2.

Rahman, N. A., Bakar, Z. A. and Sembok, T. M. T. (2010) ‘Query Expansion using Thesaurus in Improving Malay Hadith Retrieval System’, pp. 1404–1409.

Rasyidi, I., Romadhony, A. and Wibowo, A. T. (2013) ‘Indonesian Hadith Retrieval System using Thesaurus’, pp. 285–288.

Ryansyah, A. and Andayani, S. (2017) ‘Implementasi Algoritma TF-IDF Pada Pengukuran Kesamaan Dokumen’, *Jurnal Sistem & Teknologi Informasi Komunikasi*, 1(1), pp. 1–10.

Saneifar, H. *et al*. (2014) ‘Enhancing passage retrieval in log files by query expansion based on explicit and pseudo relevance feedback’, *Computers in Industry*, 65(6), pp. 937–951. doi: 10.1016/j.compind.2014.02.010.

Soelistyo, H. (2011) ‘Plagiarisme : Pelanggaran Hak Cipta dan Etika’, *Plagiarisme : Pelanggaran Hak Cipta dan Etika*, (Yogyakarta: Kanisius), p. No Pages. Available at: http://www.dt.co.kr/contents.html?article_no=2012071302010531749001.

Stein, B., Eissen, S. M. zu and Potthast, M. (2007) ‘Strategies for retrieving plagiarized documents’, *Kidney and Blood Pressure Research*, 24(2), pp. 84–91. doi: 10.1159/000054212.

Stein, B. and zu Eissen, S. M. (2006) ‘Near Similarity Search and Plagiarism Analysis’, (1993), pp. 430–437. doi: 10.1007/3-540-31314-1_52.

Vijayarani, S., Ilamathi, J. and Nitya (2015) ‘Preprocessing Techniques for Text Mining - An Overview’, 5(1), pp. 7–16. doi: 10.1016/j.procs.2013.05.286.