Perancangan Model *Information Retrieval* dengan Metode *Cosine Similarity* untuk Pencarian dan Persortiran Buku Teks Pelajaran *Online*

Alif Rizqi Mulyawan [1)], Salman Alfarizi [2)]
STMIK Nusa Mandiri Jakarta [1)]
alifrizqi.mulyawan@gmail.com[1)], alfarizisalman92@gmail.com [2)]

Abstract – *Textbooks generally serve as a reference in study materials from elementary school to college. Textbook lessons initially in production in the form of physical print or hardcopy but after the existence of the Internet has been widely applied in various aspects of life, textbook lessons have appeared in the form of softcopy that makes it very easy for anyone to look for it on the internet. From the various conveniences obtained there are few problems that arise, namely to make students or students confused to take the right reference that is being sought due to the very many textbooks lessons on the internet with the discussion and content is almost the same. In this article will be explained the design of textbook search engine lesson with information retrieval method using consine similarity method. Cosine similarity is a method to calculate how much resemblance between documents, by using a similarity measure function. Using this size function allows the document ranking to match the likeness or relevance of the textbooks sought.*
Keywords: information retrieval, search engine, consine similarity, textbooks lessons

Abstraksi – Buku teks pelajaran umumnya berfungsi sebagai referensi dalam bahan belajar dari mulai tingkat pendidikan sekolah dasar hingga perguruan tinggi. Buku teks pelajaran awalnya di produksi dalam bentuk cetakan fisik atau *hardcopy* tetapi setelah adanya internet sudah banyak diterapkan diberbagai aspek kehidupan, buku teks pelajaranpun sudah banyak bermunculan dalam bentuk *softcopy* yang membuat sangat mudah bagi siapapun untuk mencarinya di *internet.* Dari berbagai kemudahan yang didapat terdapat sedikit permasalahan yang timbul, yaitu membuat pelajar atau mahasiswa bingung untuk mengambil referensi yang tepat yang sedang dicarinya akibat sangat banyaknya buku teks pelajaran yang terdapat di internet dengan pembahasan dan konten yang hampir sama. Dalam artikel ini akan dijelaskan perancangan mesin pencarian buku teks pelajaran dengan metode *information retrieval* menggunakan *consine similarity method*. Cosine similarity merupakan metode untuk menghitung seberapa besar kemiripan antar dokumen, dengan menggunakan suatu fungsi ukuran kemiripan (*similarity measure*). Dengan menggunakan fungsi ukuran ini memungkinkan perangkingan dokumen sesuai dengan kemiripan atau relevan terhadap buku teks pelajaran yang dicari.
Kata Kunci: *temu balik informasi, mesin pencarian, consine similarity, buku teks pelajaran*

1. Latar Belakang

Informasi hampir menjadi sebuah kebutuhan primer dengan berjalannya perkembangan teknologi informasi. Kecepatan perubahan dan penambahan informasi menyebabkan dibutuhkannya suatu sistem yang dapat mengakses dan menyediakan berbagai informasi tersebut. Sistem diharuskan untuk dapat menyediakan kebutuhan informasi yang dibutuhkan pengguna dengan memberikan hasil yang akurat dan relevan. Permasalahannya bagaimana membuat sebuah sistem yang mampu memberikan hasil pencarian dokumen yang akurat dan relevan sesuai dengan yang diinginkan pengguna.

Kemudahan untuk mendapatkan informasi salah satunya didapat melalui internet yang mendorong pertambahan jumlah informasi digital semakin banyak dan beragam, salah satunya buku teks pelajaran sudah bertranformasi menjadi bentuk *e-book* yang awalnya buku teks pelajaran umumnya berbentuk cetakan kertas atau *hard copy*. Saat ini pengguna internet khusunya pelajar atau mahasiswa sangat mudah sekali dalam mencari buku teks pelajaran sebagai bahan acuan atau referensi dalam proses belajarnya karena sudah banyak tersedia *e-book* yang di unggah penulis-penulis serta penerbit yang beralih menulis digital. Dari kemudahan yang disediakan internet timbul sedikit permasalahan, yaitu membuat pelajar atau mahasiswa kebingungan dalam mengambil referensi yang sedang dicarinya dengan tepat akibat banyaknya *e-book* dengan pembahasan dan konten yang hampir sama. Dari permasalahan tersebut timbul pertanyaan bagaimana mencari *e-book* yang benar-benar

tepat dan akurat sesuai dengan pencarian yang dilakukan.

Salah satu metode dalam mendapatkan informasi yang akurat dan relevan yaitu dengan menggunakan sistem temu balik informasi *(information retrieval)*, sistem ini membuat perhitungan untuk menentukan apakah sebuah informasi relevan dengan kebutuhan penggunanya.

Dalam artikel ini akan dibangun suatu perancangan model pencarian temu balik informasi *(information retrieval)* untuk buku teks pelajaran dengan munggunakan *Cosine Similarity.* Untuk pengumpulan data teks buku pelajaran atau *e-book* menggunakan *web crawler,* yang kemudian hasilnya disimpan kedalam basis data. Dalam menerapkan *web crawler* diharuskan mempunyai daftar *URL* yang akan di akses yang di sebut dengan istilah *seeds. Crawler* akan mengunjungi daftar *URL* tersebut, di lanjutkan dengan mengidentifikasi semua *hyperlink* dari *page* tersebut dan menambahkan kembali kedalam *seeds* yang sering disebut dengan proses *crawler frontier.* Proses akhirnya *web crawler* membawa data-data yang dicari *user* kemudian menyimpan kedalam *storage.* Tugas *cosine similarity* menghitung kemiripan *query* yang dimasukan, dengan isi dokumen dan dilakukan perangkaian. Hasilnya membentuk data buku teks pelajaran atau *e-book* yang relevan satu sama lain yang dibutuhkan oleh pengguna berdasarkan kata kunci yang dicarinya. Dan pada akhirnya dapat memudahkan pelajar atau mahasiswa untuk memperoleh buku referensi yang akurat dan relevan sesuai dengan pencarian yang dilakukannya.

2. Referensi

2.1 Information retrieval

Information retrieval (IR) merupakan bidang keilmuan yang mempelajari cara-cara penelusuran kembali atas dokumen-dokumen yang ada dalam basis data dengan melakukan suatu pencarian atas dokumen-dokumen yang diinginkan pengguna, dengan melihat tingkat kemiripannya (Kurniawan, Solihin, & Hastarita, 2014).

Data yang ditelurusi bisa berupa teks, tabel, gambar (image), video, audio merupakan jenis data yang sering dipakai dalam proses pencarian. Salah satunya dengan mengurangi dokumen pencarian yang tidak relevean atau meretrieve dokumen yang relevan yang bertujuan untuk memenuhi informasi pengguna sebagai acuan dalam melakukan panggilan *(searching),* index *(indexing),* pemanggilan data kembali *(recalling)* (Sani, Zeniarja, & Luthfiarta, 2016).

Dengan demikian Information Retrieval merupakan teknik pencarian data yang dinamis artinya dapat menggunakan berbagai macam metode dengan tujuan dapat memenuhi informasi yang di inginkan pengguna. Dan data yang akan dicari tidak hanya berbentuk dokumen saja tetapi bisa berbagai data seperti multimedia dengan syarat dapat tersimpan di dalam database sebagai tempat penyimpanannya.

2.2 Cosine Similarity

Dalam Information Retrieval banyak metode yang dapat diterapkan salah satunya adalah metode Cosine Similarity. Metode ini dikhususkan untuk menghitung tingkat kesamaan antar dua buah objek. Metode cosine similarity ini menghitung tingkat kesamaan antara dua buah objek (misalkan D1 dan D2) yang dinyatakan dalam dua buah *vector* dengan menggunakan *keywords* (kata kunci) dari sebuah dokumen sebagai ukuran (Nurdiana, Jumadi, & Nursantika, 2016).

Cosine similarity dapat di katakan salah satu metode yang dapat digunakan dalam pencarian data dengan cara menghitung tingkat kesamaan antara dua buah objek, contoh: d1 dengan d2, dimana d1 merupakan kumpulan buku teks pelajaran online dan d2 adalah keyword yang hendak dicari.

2.3 TF-IDF

Agar metode cosine similarity dapat diterapkan dengan optimal pada kasus pencarian buku teks pelajaran online ini, sebelum di hitung tingkat kesamaan antara buku teks pelajaran online yang satu dengan yang lainnya sesuai dengan kata kunci yang dicari di lakukan pembobotan kata pada masing-masing buku teks pelajaran dengan menggunakan metode TF-IDF.

Term Frequency dan Inverse Document Frequency (TF-IDF) merupakan pembobotan diperuntukan untuk penelusuran informasi. Term frequency (TF) adalah pembobotan yang sederhana dimana penting atau tidaknya sebuah kata dianggap sama atau sebanding dengan jumlah kemunculan kata tersebut dalam dokumen,

sedangkan inverse document frequency (IDF) adalah pembobotan yang mengukur penting sebuah kata dalam dokumen dilihat pada seluruh dokumen secara global (Purwanti, Science, & Kesehatan, 2015).

TF-IDF dapat di artikan juga perhitungan bobot tiap kata yang di cari di tiap dokumen yang menjadi sumber pencarian serta dilakukan perankingan dokumen yang dianggap paling relevan terhadapat kata kunci yang sedang di cari

2.4 Web Crawler

Untuk mengumpulkan website atau *link* target dimana buku teks pelajaran berada digunakan Web Crawler. Web crawler merupakan *software* yang digunakan untuk menelusuri serta mengumpulkan halaman-halaman web yang selanjutnya diindeks oleh mesin pencari. Sedangkan proses crawling sendiri adalah proses yang digunakan oleh mesin pencari (search engine) untuk mengumpulkan halaman website (Aditya, 2015).

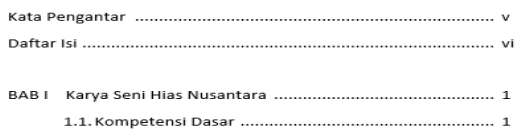
3. Metode Penelitian

3.1 Populasi Data

Dalam pembuatan model information retrieval yang penulis buat dengan studi kasus pencarian buku teks pelajaran online untuk populasi data bersumber dari situs atau web yang menyediakan konten buku teks pelajaran dengan gratis, salah satu contohnya web kemdikbud yang menyediakan e-book dari tinggal SD sampai dengan SLTA sederajat. Populasi data sendiri merupakan kumpulan data yang menjadi target dari sebuah pengujian atau riset.

3.2 Format dan Karakteristik Data

Setiap data pasti mempunyai karakteristik dan format yang berbeda tergantung apa yang sedang di uji datanya atau data yang akan di implementasikan kedalam sebuah model sistem. Studi kasus kali ini mengambil data buku teks pelajaran online (e-book) yang mempunyai format serta karakterisktik terstruktur dengan alur bab dan sub bab dan yang membedakan data satu dan lainnya hanya pada jumlah dari Bab dan Sub Bab. Berikut contoh format data buku teks pelajaran online yang di gambarkan pada daftar isi:

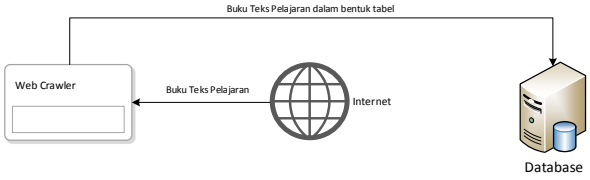

Kata Pengantar v
Daftar Isi vi

BAB I Karya Seni Hias Nusantara 1
1.1. Kompetensi Dasar 1

Gambar 1. Karakteristik Data

3.3 Organisasi Penyimpanan Data atau Informasi

Perancangan model information retrieval yang akan dibuat menggunakan web crawler yang akan melakukan penjelajahan menyeluruh pada semua halaman web yang menjadi terget data. Kemudian hasil penjelajahan tersebut akan di kumpulkan dan di simpan kedalam database sebelum data tersebut di proses oleh algoritma yang akan di pakai dan mengeluarkan hasilnya. Database yang digunakan dalam rancangan model information retrieval yang akan dibuat menggunakan mysql dengan server local pada komputer penulis. Berikut gambar struktur organisai penyimpananannya:

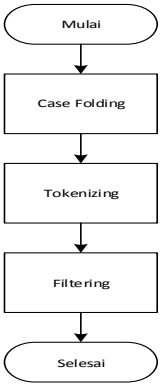


Gambar 2. Rancangan Penyimpanan Data

3.4 Preprocessing

Merupakan sebuah proses awal dalam membentuk kata kunci atau keyword-keyword dari kumpulan buku teks pelajaran online yang nantinya akan disimpan kedalam database, dengan beberapa tahapan proses didalamnya:

1. Case folding, merupakan awal proses dimana semua huruf dalam dokumen di rubah menjadi huruf kecil. Proses ini hanya menerima huruf ‘a’ sampai dengan huruf ‘z’ saja dan karakter-karakter serta tanda baca lainnya selain huruf dihilangkan.
2. Tokenizing, merupakan proses setelah dilakukan case folding dimana dilakukan pemenggalan tiap kata pada isi buku teks pelajaran berdasarkan spasi dan tanda baca penghubung seperti ‘-‘.
3. Filtering, proses terakhir dari prepocessing yaitu proses penghilangan kata-kata yang dianggap sebagai kata yang jarang dicari atau jarang digunakan dalam menentukan keyword pencarian. Ini bertujuan untuk mengurangi waktu yang dibutuhkan saat perhitungan frekuensi yang akan muncul tiap kata.



Gambar 3. Flowchart proses preprocessing

3.5 Sintaks Pencarian Informasi

Model information retrieval yang akan di buat seperti yang di jelaskan pada sub bab sebelumnya menggunakan bantuan software web crawler. Web crawler ini juga di terapkan oleh mesin pencarian terbesar di dunia google dalam proses pencariannya. Dalam pencarian informasi pengguna cukup memasukan kata kunci yang ingin dicari. Misal pengguna ingin mencari teori dari sebuah gerhana bulan, pengguna cukup mengetikan "Gerhana bulan" maka model search engine akan menampilkan beberapa hasil penyaringan dari beberapa sumber buku teks pelajaran online yang mempunyai pemnahasan tentang gerhana bulan.

3.6 Algoritma Pencarian

Secara garis besar untuk penjelasan struktur pencarian informasi pada model yang di rancang mempunyai urutan di mulai dari informasi yang di cari menggunakan web crawler kemudian perhitungan tingkat kesamaan menggunakan algoritma cosine similarity, hasilnya di proses kembali untuk di indeks agar hasil lebih mengerucut atau tingkat akurasi yang dihasilkan lebih besar menggunakan TF – IDF, terakhir menampilkan hasil yang di cari oleh pengguna.

3.7 Akurasi Pencarian

Hasil pencarian yang dihasilkan oleh perancangan model information retrieval yang dibuat mempunyai tingkat akurasi yang cukup tinggi dengan hasil pencarian yang sesuai dengan kata kunci yang di maksud oleh pengguna, ini berkat kombinasi dari algoritma cosine similarity dengan metode pengindeksan TF – IDF. Berikut simulasi contoh dari pencarian menggunakan metode diatas:

D1: Tokoh politik dari berbagai partai mengadakan untuk membahas koalisi baru.
D2: Partai demokrat memenangkan pemilu 2009 karena figur pak SBY.
D3: Pertandingan pertama antara Persib dan Persija diadakan di jakarta.
D4: Beberapa pertandingan sepak bola yang dilakoni Persebaya medapat hasil lebih banyak menang.

Q: "Menang Pertandingan"

Jika yang dimaksud pengguna adalah pencarian kemenangan partai maka tidak bisa hanya memakai kata kunci "menang pertandingan" saja karena yang akan ditampilkan untuk pengindeksan pertama adalah informasi tentang bola. Jadi semakin lengkap kata kunci semakin besar tingkat akurasi pencariannya.

3.8 Evaluasi

Evaluasi digunakan untuk mengukur kinerja dari suatu sistem untuk menghasilkan perbaikan pada proses pengambilan informasi. Ukuran yang menyatakan pengukuran kualitas dari text retrieval umumnya menggunakan recall, precision dan f-measure.

Recall, adalah hasil perhitungan dari proporsi jumlah dokumen teks yang revelan terkenali di antara semua dokumen teks relevan yang ada pada koleksi (tersimpan dalam database).

$$Recall = \frac{Relevan \cap Retrieved\ Buku\ Teks\ Pelajaran\ Online}{Relevan\ Buku\ Teks\ Pelajaran\ Online} \ldots\ldots$$

Precision, adalah hasil perhitungan proporsi jumlah dokumen teks yang relevan terkenali di antara semua dokumen yang terpilih oleh sistem.

$$Precision = \frac{Relevan \cap Retrieved\ Buku\ Teks\ Pelajaran\ Online}{Retrieved\ Buku\ Teks\ Pelajaran\ Online} \ldots\ldots$$

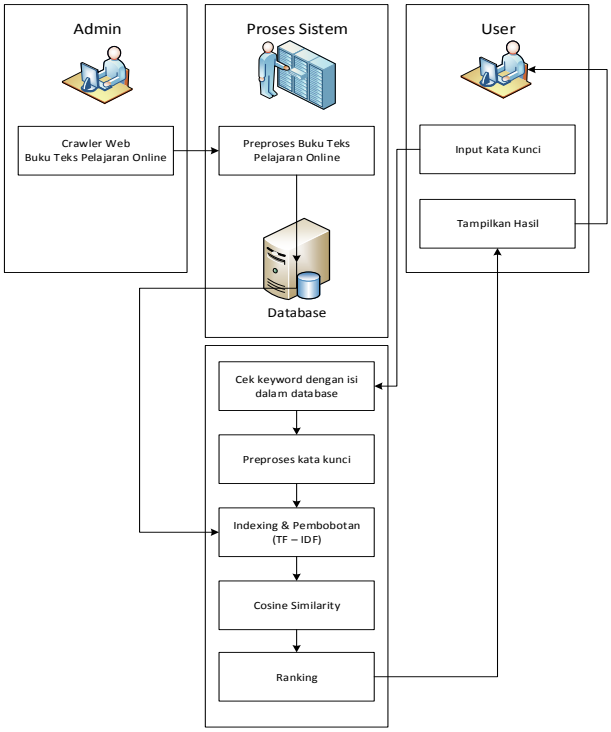
F-Measure, adalah hasil akhir berbentuk nilai yang mewakili seluruh kinerja sistem dan merupakan rata-rata dari nilai precision dan recall.

$$F\text{-}Measure = \frac{2\ Precision\ x\ Recall}{Precision + Recall} \ldots\ldots\ldots$$

4. Hasil dan Pembahasan

4.1 Perancangan

Sebelum membuat program aplikasi, terlebih dahulu dilakukan proses perancangan sistem. Hal ini dilakukan agar aplikasi yang dibuat dapat berfungsi sesuai dengan yang di harapkan dan maksimal. Gambaran umum dari sistem ini dapat dilihat pada gambar 3.

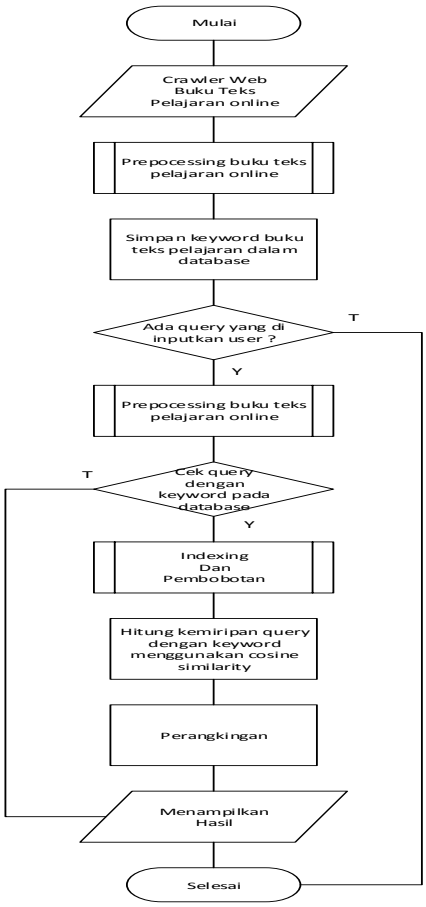


Gambar 4. Gambaran umum sistem

Gambar diatas dapat di deskrispsikan sebagai berikut. Yang pertama admin, merupakan pengguna yang berhak melakukan kontrol penuh dari sistem dan database. Admin mengcrawler web teks buku pelajaran online lalu disimpan kedalam database. Kemudian sistem melakukan proses preproses pada data buku teks pelajaran online, sehingga dihasilkan keyword buku teks pelajaran yang tersimpan dalam database dilakukan indexing dan pembobotan.

User adalah pengguna yang akan menginput query dan melihat informasi hasil pencarian. User menginput query kedalam aplikasi, kemudian sistem melakukan proses preproses query sehingga diperoleh keyword query. Selanjutnya sistem akan melakukan pecocokan keyword query dengan keyword buku teks pelajaran yang telah tersimpan dalam basis data. Jika ada keyword yang cocok atau sama, maka keyword query akan di index dan dilakukan proses pembobotan. Tapi jika tidak ada yang cocok, maka tidak ada hasil yang ditampilkan. Setelah proses indexing dan pembobotan kemudian dilakukan proses perhitungan kemiripan cosine similarity antara bobot keyword buku teks pelajaran online dan bobot keyword query, lalu dilakukan proses perangkingan. Dan terakhir adalah mengukur kinerja sistem, dilakukan evaluasi menggunakan recall, precision dan f-measure.

1. Flowchart Sistem



Gambar 5. Flowchart Sistem

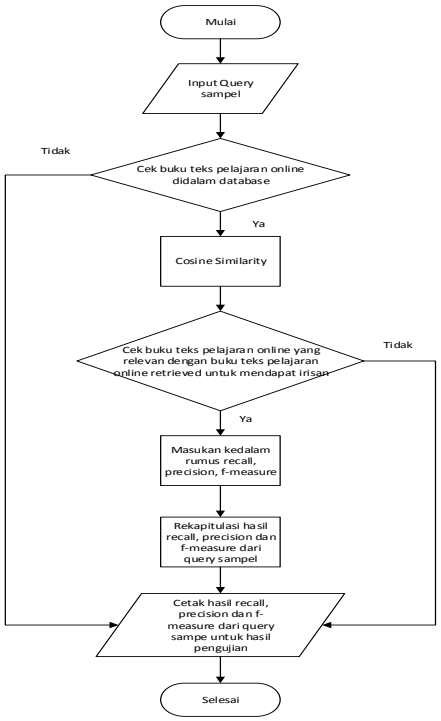
Deskripsi keterangan dari flowchart diatas dapat dijelaskan sebagai berikut:

1. Model sistem pecarian melakukan crawler pada website target tempat banyak buku teks pelajaran online yang akan dicari seperti https://belajar.kemdikbud.go.id/
2. Simpan semua artikel atau buku teks pelajaran online kedalam database.
3. User memasukan kata kunci atau query yang ingin dicari dalam pencarian buku teks pelajaran online.
4. Cek query yang di input oleh user kedalam database yang sudah berisi banyak buku teks pelajaran.
5. Kemudian sistem melakukan indeks dan pembobotan pada query yang di cari oleh user.
6. Dari hasil indeks dan pembobotan maka dilakukan perhitungan kemiripan antara query dan keyword menggunakan algoritma cosine similarity.
7. Maka akan banyak hasil yang akan didapatkan, untuk lebih meningkatkan akurasi pencarian di lakukan perangkingan.

8. Maka hasil akan ditampilkan.

4.2 Skenario Uji Coba Sistem

Uji coba yang dilakukan adalah dengan melakukan perbandingan nilai terkecil dari cosine similarity yang dapat di retrieve oleh model sistem yang di buat dengan proses pengujian sebagai berikut:



Gambar 6. Flowchart Pengujian Sistem

Flowchart diatas bisa di deskripsikan sebagai berikut:

1. Pengujian model dilakukan menggunakan 15 query sampel dengan rincian 5 query yang terdiri dari 1 kata, 5 query terdiri dari 2 kata dan 5 query terdiri dari 3 kata.
2. Agar mendapat nilai yang relevant buku teks pelajaran online, masing-masing query di cek pada keselurahan buku teks pelajaran online yang ada didalam database yang mengandung kata query.
3. Jika data terdapat yang cocok, maka dilakukan pembatasan nilai terkecil cosine similarity.
4. Selanjutnya relevan buku teks pelajaran online di cek dengan buku teks pelajaran online yang teretrieve guna mendapatkan irisan relevan dari retrieve buku teks pelajaran online.
5. Jika terdapat data yang sama kembali akan dilakukan evaluasi dengan mencari nilai recall, precision dan f-measure.

5. Kesimpulan

Bedasarkan hasil dari perancangan model information retrieval menggunakan cosine similarity yang digunakan untuk pencarian buku teks pelajaran online dapat ditarik kesimpulan model harus mampu mengumpulkan buku teks pelajaran online melalui fasilitas web crawler dan memberikan bobot dengan mengimplementasikan pembobotan menggunakan metode TF-IDF dan penyeleksian query dengan algoritma cosine similaruty yang akan membuat data lebih banyak relevan di perhitungkan dalam pencarian yang dilakukan.

6. DAFTAR PUSTAKA

[1] Aditya, B. R. (n.d.). Penggunaan Web Crawler Untuk *Menghimpun Tweets dengan Metode Pre-Processing Text Mining*.

[2] Belakang, A. L., Muhammad, N., & Al-quran, S. A. W. (2016). *PERBANDINGAN METODE COSINE SIMILARITY DENGAN METODE JACCARD SIMILARITY PADA APLIKASI PENCARIAN TERJEMAH AL-QUR ' AN*, *I*(1), 59–63.

[3] Komputer, J. I. (2012). Jurnal Ilmu Komputer - Volume 5 - No 2 – September 2012, *5*(2).

[4] Kurniawan, A., Solihin, F., & Hastarita, F. (2014). *PERANCANGAN DAN PEMBUATAN APLIKASI PENCARIAN INFORMASI BEASISWA DENGAN MENGGUNAKAN COSINE SIMILARITY*, *4*(2), 115–124.

[5] Purwanti, E., Science, L., & Kesehatan, I. (2015). *Klasifikafi Dokumen Temu Kembali Informasi dengan K-Nearest Neghbour Information Retrieval Document Classified with K-Nearest Neighbor*, *1*, 129–138.

[6] Report, T. (n.d.). *Penerapan Algoritma K-Nearest Neighbor pada Information Retrieval dalam Penentuan Topik Referensi Tugas Akhir*, *1*(2), 123–133.