APLIKASI PENCARIAN PERGURUAN TINGGI DENGAN ALGORITMA K-NEAREST NEIGHBOR BERBASIS INFORMATION RETRIEVAL DAN GEOGRAPHIC INFORMATION SYSTEM

Junta Zeniarja[1], Ardytha Luthfiarta[2], Catur Supriyanto[3]
[1,2,3]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
e-mail: [1]junta@dsn.dinus.ac.id, [2]ardytha.luthfiarta@dsn.dinus.ac.id, [3]catur.supriyanto@dsn.dinus.ac.id

ABSTRAK

Informasi tentang letak geografis perguruan tinggi memang sangat diperlukan bagi lulusan pelajar Sekolah Menengah Atas (SMA) yang ingin melanjutkan pendidikannya ke jenjang yang lebih tinggi. Oleh karena itu, perlu adanya aplikasi Geographic Information Systems (GIS) berbasis Information Retrieval (IR) untuk peningkatan pencarian informasi lokasi perguruan tinggi, yang diharapkan dapat mempermudah calon mahasiswa baru untuk mengetahui letak secara pasti beberapa perguruan tinggi. Di sisi lain, juga dapat menjadi sarana untuk mempromosikan perguruan tinggi yang semakin menjanjikan. Dalam penelitian ini, menggunakan sampling data beberapa perguruan tinggi yang berada di Semarang yang diperoleh dari studi pustaka dan survei lapangan. Aplikasi GIS berbasis IR ini akan dibangun dengan menggunakan pemrograman web. Dimana web tersebut berfungsi sebagai mesin pencari ketika seseorang ingin mencari perguruan tinggi, maka akan muncul map dan informasi perguruan tinggi yang berada di daerah tersebut. Map di representasikan dengan text kategorial dan query tersebut di cari menggunakan metode klasifikasi K-Nearest Neighbor (K-NN). Hasil dari penelitian ini berupa model penerapan algoritma K-Nearest Neighbor yang tepat dalam hal pencarian query yang sesuai di dalam dokumen corpus dan menggunakan API Google Maps yang akan menampilkan peta hasil query berdasarkan kemiripan (similarity) antara query tersebut dengan kategori teks.

Kata Kunci*: perguruan tinggi, SMA, GIS, IR, K-NN*

1. PENDAHULUAN

Perkembangan internet telah hadir dan membawa perubahan dalam kehidupan manusia. Hampir setiap orang dalam melakukan aktivitasnya tidak terlepas dengan internet. Pada tahun 2014, di Indonesia terdapat 83,7 orang yang aktif menggunakan internet sehingga Indonesia menjadi peringkat ke-6 terbesar di dunia, berdasarkan info dari lembaga riset pasar e-Marketer. Kemudian pada tahun 2015, secara menyeluruh mencapai 3 miliar orang jumlah pengguna internet di seluruh dunia. Dan pada akhirnya nanti, pada tahun 2018, pengguna yang mengakses internet setidaknya sekali tiap satu bulan diperkirakan sebanyak 3,6 miliar orang di bumi[1].

Berdasarkan data statistik pengguna Internet diatas, diketahui bahwa internet memiliki pengaruh yang sangat kuat dalam kehidupan manusia dan telah menjadi bagian hidup dari manusia di dunia. Salah satu manfaat dari penggunaan internet dalam bidang ilmu pengetahuan adalah untuk mengakses dan mencari berbagai macam informasi. Sebagian besar orang mencari informasi tentang suatu permasalahan melalui internet dengan bantuan mesin pencari seperti Google. Terutama bagi pelajar lulusan Sekolah Menengah Atas (SMA) yang ingin mendapatkan informasi mengenai perguruan tinggi baik perguruan tinggi negeri (PTN) maupun perguruan tinggi swasta (PTS) melalui internet. Menurut hasil pengamatan dari Direktorat Jenderal Pendidikan Tinggi (DIKTI), pada tahun 2012 di Indonesia terdapat 3150 perguruan tinggi baik PTN maupun PTS serta 15.830 program studi [2]. Sedangkan sekarang ini terdapat 4.438 perguruan tinggi yang terdiri atas 1.105 akademi, 241 politeknik, 2.421 sekolah tinggi, 130 institut, dan 541 universitas [3]. Sehingga ruang lingkup penelitian perguruan tinggi di Indonesia cukup luas berdasarkan fakta tersebut. Jumlah persebaran perguruan tinggi tersebut juga menjadi potensi dan tantangan bagi pemerintah untuk menyelenggarakan pendidikan yang adil dan merata bagi rakyat Indonesia.

Informasi yang akurat tentang PTN/PTS di Indonesia secara spesifik masih jarang media yang menyediakan. Oleh karena itu, banyak pelajar lulusan SMA merasa kesusahan dalam mencari informasi yang tepat dan terbaru tentang PTN/PTS di Indonesia. Untuk memudahkan para pelajar dalam pencarian lokasi PTN/PTS secara detil, diperlukan sebuah aplikasi pemetaan yang dapat diakses melalui internet yang dapat menunjukkan lokasi secara detail PTN/PTS yang hendak mereka tuju secara visual.

Dalam hal pencarian, erat kaitannya dengan konsep Information Retrieval (IR) atau lebih dikenal dengan nama temu kembali informasi. Ilmu yang mempelajari metode dan prosedur untuk menemukan kembali informasi yang tersimpan dari berbagai sumber yang relevan maupun koleksi sumber informasi yang dicari serta dibutuhkan merupakan konsep dasar dari IR. Dalam pencarian data berupa dokumen, beberapa jenis dokumen dapat ditemukan diantaranya berupa teks, tabel, gambar, video, dan audio. Untuk mencukupi kebutuhan informasi pengguna dengan cara meretrieve dokumen yang relevan atau mengurangi pencarian dokumen yang tidak relevean merupakan tujuan yang hendak dicapai oleh IR.

Beberapa penelitian yang mendiskusikan permasalahan tersebut antara lain pendekatan supervised learning dengan klasifikasi atau kategorisasi teks yang saat ini mempunyai banyak cara pendekatannya seperti berbasis numeris, misalnya pendekatan probabilistic, Support Vector Machine (SVM), Artificial Neural Network

(NN), K-Nearest Neighbor (KNN), dan juga berbasis non numeris diantaranya Decision Tree (DT). Dalam pendekatan KNN yang berbasis numeris memiliki beberapa keunggulan yaitu: cepat, berakurasi tinggi dan sederhana. Dalam algoritma KNN, jarak atau perbedaan atribut kata yang hadir pada suatu dokumen menjadi dasar pengklasifikasian dokumen teks. Kinerja KNN sebagai algoritma klasifikasi cukup bagus, dibuktikan oleh beberapa penelitian yang menggunakannya [4].

Aplikasi Sistem Informasi Geografis atau dikenal dengan nama Geographic Information Systems (GIS) dapat dibangun menggunakan konsep IR melalui metode kategori teks dengan algoritma KNN, dimana dengan membuat teks kategori untuk maps memungkinkan untuk memudahkan dalam mencari lokasi perguruan tinggi. Misalnya dengan mengkategorikan perguruan tinggi berdasarkan letak geografisnya, jenis perguruan tingginya dan pengkategorian lainnya. Dengan pengkategorian teks di dalam IR menggunakan metode K-Nearest Neighbor yang diterapkan ke dalam aplikasi GIS untuk pencarian informasi perguruan tinggi, diharapkan para pelajar yang ingin melanjutkan pendidikan ke perguruan tinggi akan dimudahkan dalam pencarian informasi maupun lokasinya hanya dengan berbekal informasi minimal yang mereka ketahui.

2. TINJAUAN PUSTAKA

3.1. Tinjauan Studi

Beberapa studi yang telah dilakukan oleh peneliti sebelumnya, hasil yang didapatkan menunjukkan berbagai tinjauan tentang penerapan konsep Information Retrieval ke dalam aplikasi GIS (Geographic Information System). Penelitian sebelumnya yang relevan dengan penelitian ini antara lain sebagai berikut:

a. User Profiling for University Recommender System Using Automatic Information Retrieval [5].

Dalam penelitian ini, mereka membahas mengenai profiling pengguna untuk sistem rekomendasi universitas melalui proses ekstraksi, mengintegrasikan dan mengidentifikasi informasi berdasarkan kata kunci untuk menghasilkan profil yang terstruktur dan kemudian memvisualisasikan pengetahuan dari temuan ini. Pengguna profiling membantu personalisasi sistem yang bekerja sesuai dengan pengguna. Oleh karena itu pengguna profil atau personalisasi yang digunakan untuk mengakses informasi pengguna yang relevan, yang dapat digunakan untuk memecahkan masalah yang sulit dari sistem rekomendasi seperti klasifikasi dan peringkat item sesuai dengan kepentingan individu. Mereka menggunakan model ekstraksi profil untuk mengambil data dari berbagai sumber web dan berdasarkan situs jejaring sosial mereka.

b. Architecture of a concept-based information retrieval system for educational resources [6].

Dalam penelitiannya yang mereka usulkan, Perez, Anido, Gomez dan Maurino menyajikan SDE: Search (pencari), Discovery (menemukan) dan Explore (menjelajahi), sebuah mesin pencari eksplorasi sumber daya berbasis pendidikan yang dibangun di atas pengetahuan yang diberikan oleh Wikipedia: sekumpulan artikel yang memberikan ruang pencarian (set topik yang pengguna dapat menyelidiki), dan hubungan antara artikel Wikipedia menginformasikan saran bahwa mesin pencari memberikan kepada siswa untuk masuk lebih dalam eksplorasi domain tertentu tentang pengetahuan. SDE indeks beberapa ratusan ribu sumber daya pendidikan dari sumber Web berkualitas tinggi, seperti Project Gutenberg dan Pendidikan Terbuka di Eropa, dan masih banyak lagi. Selain itu mereka juga melaporkan hasil evaluasi SDE oleh para ahli di bidang Teknologi Pembelajaran Peningkatan di beberapa lokakarya yang berlangsung di seluruh Eropa dalam konteks ITEC proyek Eropa FP7. Hasil ini memungkinkan mereka untuk menyimpulkan bahwa paradigma pencarian eksplorasi, memanfaatkan pengetahuan yang diperoleh dari Wikipedia, adalah pendekatan yang sangat menjanjikan untuk membangun sistem pencarian informasi untuk digunakan dalam konteks pembelajaran.

c. Geographic information retrieval: Modeling uncertainty of user's context [7].

Pada papernya telah dijelaskan bahwa, Bordogna, Ghisalberti dan Psaila mengusulkan suatu model Geographic Information Retrieval dan sistem mengimplementasikannya yang mewakili kedua ketidakpastian dalam mengindeks konten dokumen geografis dan konteks pengguna serta preferensi dalam mengevaluasi query spasial yang fleksibel. Mengekstrak konten geografis dari dokumen teks dan menerapkan pengetahuan kode heuristik oleh aturan bipolar yang mengevaluasi petunjuk positif dan petunjuk negatif bagi pengakuan nama geografis di dalam teks. Dengan demikian, itu merupakan konten geografis dokumen oleh fuzzy, yaitu, lokasi yang berbeda di bumi yang terkait dengan teks dengan tingkat yang berbeda dari signifikansinya. Akhirnya, sistem memungkinkan mengevaluasi dua jenis query fleksibel menggabungkan kondisi berdasarkan konten dengan kondisi tata ruang.

d. Evaluation on geospatial information extraction and retrieval: Mining thematic maps from web source [8].

Dalam penelitiannya, mereka memanfaatkan informasi geospasial web. Survei ini dilakukan dalam konteks metode-metode ekstraksi, pengambilan, visualisasi, dan kemungkinan skenario pertambangan atau penemuan pengetahuan lebih lanjut untuk menghasilkan peta tematik secara otomatis dari korpus web. Mereka menemukan bahwa Geographic Information Retrieval (GIR) berbasis Web, metode yang mengembalikan bidang terpilih yang relevan tetapi masih terdapat beberapa poin yang kurang, meskipun pemodelan area umumnya di GIS. Sebagian besar metode GIR masih difokuskan pada tempat-tempat dan bangunan, bukan tema atau informasi sekitar beberapa daerah. Oleh karena itu, menunjukkan bahwa keadaan metode GIR belum cukup untuk ekstraksi tematik dan pengambilan untuk menghasilkan peta tematik dari korpus web menggunakan model

topik Bayesian seperti *Latent Dirichlet Allocation* dapat berfungsi sebagai dasar yang baik untuk melayani kasus penggunaan tersebut.

e. Application of Text Summarization techniques to the Geographical Information Retrieval task [9].

Di dalam papernya menjelaskan tentang konsep Peringkasan Teks secara otomatis yang telah terbukti berguna untuk tugas-tugas pemrosesan bahasa alami (*Natural Language Processing*) seperti *Question Answering* atau Klasifikasi Teks dan bidang terkait lainnya dari ilmu komputer seperti penanganan Informasi Retrieval. Mereka juga membahas mengenai konsep Geografis Information Retrieval yang dapat dianggap sebagai perpanjangan dari bidang Information Retrieval, generasi ringkasan dapat diintegrasikan ke dalam sistem ini dengan bertindak sebagai tahap peralihan, dengan tujuan mengurangi panjang dokumen. Dengan cara ini, waktu akses untuk mencari informasi akan ditingkatkan, sementara pada saat yang sama dokumen yang relevan akan juga diambil. Oleh karena itu, mereka mengusulkan generasi dua jenis ringkasan (generik dan grafis geografis) yaitu menerapkan beberapa tingkat kompresi untuk mengevaluasi efektivitasnya didalam konsep Geografis Information Retrieval.

f. Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering [10].

Di dalam penelitian yang telah mereka lakukan, Nesi, Pantaleo, dan Tenti membahas masalah mengumpulkan informasi geografis dari teks yang tidak terstruktur di halaman web maupun dokumen. Di dalam spesifik, metode yang diusulkan bertujuan untuk penggalian lokasi geografis (di resolusi nomor jalan) perusahaan dan jasa komersial, oleh annotating informasi geografis yang terkait dari domain web mereka. Proses penjelasan berdasarkan pemrosesan bahasa alami (*Natural Language Processing*) teknik untuk teks komprehensif, dan bergantung pada *Pattern Matching* dan pengakuan *Hierarchical Cluster Analysis* dan entitas geografis yang *dis-ambiguating*. Hasil dari performansi Geotagging telah dinilai dengan mengevaluasi Presisi, Recall dan F-Measure output sistem yang diusulkan (diwakili dalam bentuk semantik RDF tiga kali lipat) terhadap kedua database referensi *geo-annotated* dan repositori semantik dari Smart City.

3.2. Landasan Teori

a. Information Retrival (IR)

Information Retrieval juga dikenal dengan nama Temu Kembali Informasi yaitu ilmu di dalam pencarian informasi yang terdapat pada dokumen, pencarian untuk metadata yang menjelaskan dokumen, pencarian untuk dokumen itu sendiri, atau pencarian di dalam database, baik relasi database yang stand-alone atau hipertext database yang terdapat pada network seperti internet maupun world wide web atau intranet, berupa teks, suara, gambar, atau data. Sesuai dengan prinsipnya bahwa penyimpanan informasi dan penemuan kembali informasi adalah hal yang sederhana. Misalkan terdapat tempat penyimpanan banyak dokumen dan pengguna (user) merumuskan suatu pertanyaan berupa *request* atau *query* yang jawabannya adalah himpunan dokumen yang mengandung informasi yang diperlukan yang diekspresikan melalui *keyword* user. User bisa saja memperoleh dokumen-dokumen yang diperlukannya dengan membaca semua dokumen di dalam tempat penyimpanan, menyimpan dokumen-dokumen yang relevan dan membuang dokumen lainnya yang kurang relevan.

Hal ini merupakan perfect retrieval, tetapi solusi ini tidak praktis. Karena user tidak memiliki waktu atau tidak ingin menghabiskan waktunya untuk membaca seluruh koleksi dokumen, terlepas dari kenyataan bahwa secara fisik user tidak mungkin dapat melakukannya. Oleh karena itu, diperlukan suatu sistem temu kembali informasi (Information Retrieval System) untuk membantu user menemukan dokumen yang diperlukannya. Information Retrieval System merupakan sistem dengan menerapkan fungsi kesamaan dengan peringkat satu set item (biasanya dokumen tekstual) dalam menanggapi permintaan pengguna [11].

b. Geography Information System (GIS)

Geography Information System (GIS) atau dikenal dengan nama Sistem Informasi Geografis (SIG). GIS merupakan gabungan tiga unsur pokok yaitu sistem, informasi dan geografis. Sehingga GIS merupakan suatu sistem yang menekankan pada unsur informasi geografis, dimana informasi geografis tersebut mengandung pengertian informasi tentang tempat-tempat yang berada di permukaan bumi, pengetahuan tentang letak suatu objek di permukaan bumi, dan informasi tentang keterangan-keterangan (atribut) yang terdapat di permukaan bumi yang posisinya telah diketahui. Sejak akhir 1990-an GIS mulai diintegrasikan dengan sistem database spasial dalam apa yang tampaknya menjadi ide yang sempurna. Akibatnya, saat ini GIS difokuskan pada pengumpulan data spasial, mengedit, analisis dan visualisasi, sedangkan database spasial kesepakatan dengan penyimpanan data, query, pengindeksan, optimasi, dan integritas [12].

Selain itu, GIS sebagai sistem yang dirancang untuk menangkap, menyimpan, memanipulasi, menganalisis, mengelola, dan menyajikan semua jenis data geo-referenced, telah banyak digunakan sebagai alat pendukung keputusan spasial. Namun, modus interaksi satu pengguna GIS tradisional membatasi kompleksitas masalah tata ruang dipecahkan dan untuk efisiensi pemecahan masalah. Pada kenyataannya, banyak masalah tata ruang dan tugas pengambilan keputusan yang melibatkan informasi geografis memerlukan beberapa pengguna untuk bekerja sama untuk mengolah dan menganalisis data geografis. Dalam proses pengambilan keputusan dan pemecahan masalah tata ruang, ada kecenderungan mengintegrasikan real-time dan kolaborasi, yang telah menjadi salah satu daerah penting dari penelitian dan pengembangan dalam teori dan aplikasi GIS [13].

c. Geography Information Retrieval (GIR)

Geography Information Retrieval (GIR) atau lebih dikenal dengan sebutan pencarian informasi geografis merupakan spesialisasi Information Retrieval (IR) yang berhubungan dengan pengindeksan, pencarian, pengambilan dan browsing informasi bereferensi geografis, dengan penekanan pada spasial dan geografis yang berfokus pada pengindeksan dan pencarian (Bordogna et al., 2012). Mengakses informasi dengan referensi geografis dapat alami dan berguna dalam beberapa konteks: misalnya ketika mencari sumber daya di wilayah untuk tujuan wisata dan perangkat dari mobile, termasuk salah satu praktek umum dalam mengakses informasi yang relevan dengan menentukan tempat geografis yang menarik; ketika merencanakan operasi penyelamatan selama keadaan darurat yang disebabkan oleh bencana alam termasuk manfaat yang besar dalam mengambil informasi geografis, yang secara sukarela dibuat secara bebas oleh sejumlah saksi dari peristiwa tersebut. Pada umumnya, telah diperkirakan bahwa sekitar 15% dari permintaan diserahkan ke mesin pencari tujuan umum berisi nama geografis. Sehingga representasi dan Geographical Information Retrieval (GIR) saat ini menjadi trending topik penelitian saat ini. Sebagaimana didefinisikan dalam, penelitian tentang GIR menggabungkan aspek penelitian tentang IR, DBMS, user interface, GIS, dan desain sistem GIR yang efektif dan efisien. Pendekatan saat ini untuk GIR yang beragam, mulai dari IR dasar pendekatan dengan tidak ada perhatian untuk mengindeks informasi spasial dan geografis beserta penalaran, seperti mesin pencari generik lakukan, untuk lebih khusus pendekatan menerapkan tagging part-of-speech dari NLP dan pengakuan entitas nama geografis untuk mengekstraksi nama geografis dan hubungan spasial dari teks-teks maupun queries.

d. Pengkategorian Teks atau Pengklasifikasian Teks

Pengkategorian teks atau dikenal dengan nama lain pengklasifikasian teks merupakan salah satu topik di dalam Information Retrieval maupun Teks Mining yang telah memperoleh signifikan popularitas selama dekade terakhir atau lebih. Salah satu alasan utamanya adalah peningkatan jumlah dokumen digital yang tersedia dan kebutuhan untuk mengakses konten dengan cara yang lebih fleksibel. Selain itu Teks Klasifikasi juga disebut sebagai Kategorisasi Teks, Dokumen Klasifikasi, atau bahkan Topik Spotting. Pendekatan saat ini untuk teks klasifikasi menerapkan paradigma pembelajaran mesin (machine learning) yang menggunakan satu set dokumen yang sebelumnya dikategorikan untuk secara otomatis membangun categoriser dengan belajar dari data tersebut (yaitu, induktif inferensi). Sebagai bagian dari proses ini, setiap dokumen teks diwakili oleh vektor fitur, sehingga mengabaikan urutan kata-kata dan isu-isu tata bahasa lainnya, seperti representasi ini mampu menyimpan informasi yang cukup berguna untuk tugas klasifikasi [14].

e. K-Nearest Neighbor (KNN)

KNN dikenal sebagai salah satu metode berbasis NN yang paling tua dan popular di dalam melakukan pengkategorian teks. Nilai K yang digunakan merepresentasikan jumlah tetangga terdekat yang digunakan dalam menentukan prediksi label kelas pada data uji [14]. Dari K tetangga terdekat yang terpilih kemudian dilakukan voting kelas dari K tetangga terdekat tersebut. Kelas dengan jumlah suara tetangga terbanyaklah yang diberikan sebagai label kelas hasil prediksi pada data uji. Dengan data latih yang berjarak paling dekat dengan objek untuk melakukan klasifikasi dianggap sebagai metode terbaik dalam proses tersebut. Adapun cara kerja dari KNN perlu adanya penetuan inputan berupa data latih, data uji dan nilai K. Kemudian menghitung jarak data yang diuji dengan data latih dengan mengurutkan data latih berdasarkan kedekatan jaraknya. Setelah itu, pengambilan K data latih teratas untuk menentukan kelas klasifikasi untuk kelas yang dominan dari K data latih yang diambil. Dekat atau jauhnya tetangga biasanya dihitung menggunakan konsep nilai kemiripan *Cosinus Similarity*.

3. METODE PENELITIAN

Bagian ini dapat meliputi analisa, arsitektur, metode yang dipakai untuk menyelesaikan masalah, implementasi.

3.1. Instrumen Penelitian

Beberapa perangkat yang digunakan untuk mengerjakan penelitian ini adalah sebagai berikut:

a. Perangkat Lunak

Dalam melakukan penelitian, diperlukan beberapa perangkat lunak antara lain sebagai berikut:

1) Sistem Operasi : Microsoft Windows 10 Professional 64 bit.
2) Web Server : Apache.
3) Web Browser : Google Chrome, Mozilla Firefox.
4) Bahasa Pemrograman : PHP, HTML dan CSS.

b. Perangkat Keras

Beberapa perangkat keras yang diperlukan dalam melakukan penelitian adalah sebagai berikut:

1) Prosesor yang digunakan adalah Intel Core i5 2,5 GHz.
2) Memory RAM dengan ukuran 4GB.
3) Hardisk dengan ukuran 500GB.
4) Layar monitor ukuran 14 inchi.

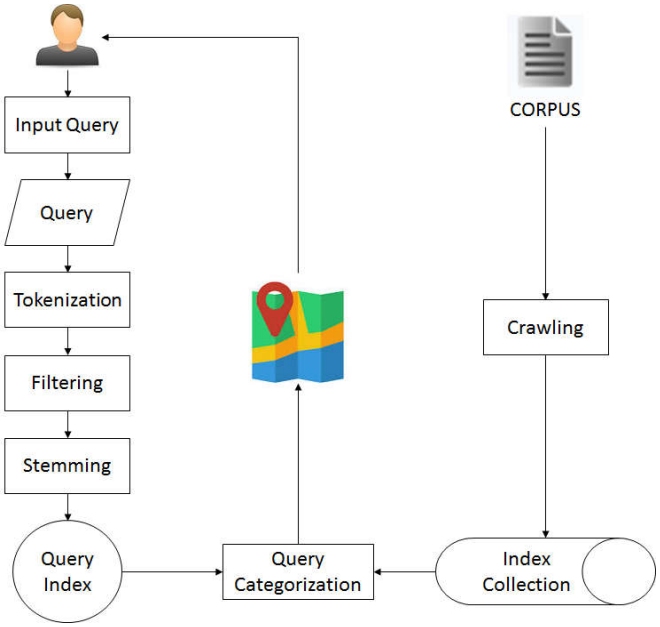
3.2. Metode Pengumpulan Data

Metode pengumpulan data yang dilakukan dalam penelitian sebagai berikut:

a. Pengambilan data dilakukan melalui dokumen elektronik atau web, dimana data yang digunakan adalah data dokumen teks Perguruan Tinggi di wilayah Semarang.
b. Pengambilan data diperoleh dari web DIKTI, BAN-PT dan survey langsung ke lokasi.
c. Pengambilan data literatur dan pendukung pustaka lainnya berasal dari paper, jurnal maupun proseding antara tahun 2012 – 2017 yang terkini.

3.3. Desain Penelitian

1) Aplikasi GIS berbasis Information Retrieval



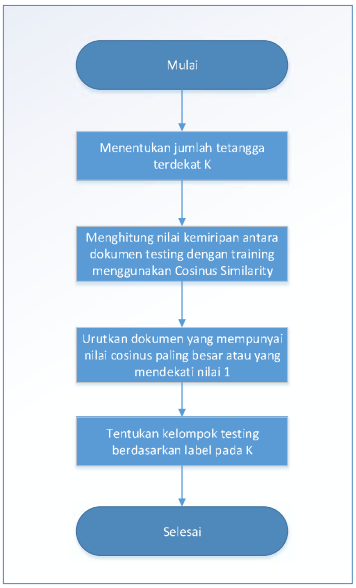
Gambar 1. Desain Penelitian aplikasi GIS berbasis Information Retrieval

Pada gambar 1 tentang desain penelitian aplikasi GIS berbasis Information Retrieval, di jelaskan beberapa tahapan dalam *preprocessing* dari pencarian teks dengan konsep Information Retrieval. Proses awal dimulai dari pengguna (user) yang akan melakukan input Query berupa informasi yang hendak dicari dan user akan mendapat output dalam bentuk peta. Operasi dari *preprocessing* dengan konsep Information Retrieval meliputi Tokenizing, Filtering, dan Steming. Di mana ketika user input Query, maka berlaku operasi pertama adalah Tokenizing yaitu proses pemecahan kata dari sebuah kalimat. Setelah melalui tokenizing, kata yang terpilih akan di lakukan operasi filtering, kata yang tidak relevan akan di buang. Kemudian, kata yang terpilih akan di berlakukan proses stemming atau pencarian kata dasar dari kata yang terpilih tadi. Kata yang sudah melewati ketiga operasi akan berkumpul menjadi sebuah index atau index Query.

Pada sistem CORPUS sendiri terjadi operasi pengumpulan data pada sistem google maps (corpus) biasa di sebut operasi crawling. Dan data yang relevan dengan Query akan masuk dalam koleksi dokumen setelah melalui tahap crawling. Output dari aplikasi ini adalah sebuah Geographical Information Retrieval dimana query yang sesuai dengan pencarian yang diinputkan oleh pengguna. Query dengan bobot tertinggi dan tingkat kemiripan tertinggi menjadi dokumen yang relevan, akan di tampilkan dalam bentuk peta (*maps*). Selain di tampilkan dalam bentuk peta, aplikasi ini juga memiliki fitur untuk melihat informasi umum dari hasi yang di tampilkan.

2) Pengkategorian Teks dengan Metode K-Nearest Neighbor

Pada gambar 2 dibawah ini, menjelaskan mengenai alur dari algoritma K-Nearest Neigbor (KNN). Dimulai dengan menentukan jumlah tetangga terdekat dari nilai K yang sudah ditentukan, kemudian menghitung nilai kemiripan dokumen testing ke dokumen training. Dokumen diurutkan berdasarkan dokumen yang mempunyai nilai kemiripan terbesar atau mendekati nilai 1, yang sebelumnya sudah dihitung terlebih dahulu menggunakan rumus Cosinus Similarity. Sehingga, dokumen tersebut akan menentukan kelompoknya berdasarkan label pada nilai K yang telah di tentukan diawal.



Gambar 2. Alur dari algoritma K-Nearest Neighbor

4. HASIL DAN PEMBAHASAN

Dari hasil penelitian ini, akan dibagi menjadi 2 tahap yaitu tahap pertama berupa tahap pembelajaran atau training yang dimulai dengan pengklasifikasian terhadap dokumen beberapa Perguruan Tinggi di Semarang yang sudah di ketahui kategorinya. Sedangkan pada tahap kedua berupa tahap pengujian atau tahap testing, hal-hal yang dilakukan adalah dengan melakukan klasifikasi dokumen beberapa Perguruan Tinggi yang belum diketahui kategorinya. Pembahasan dari hasil Penelitian yang dilakukan sebagai berikut:

a. Dokumen Teks
   Dokumen Teks yang digunakan dibagi menjadi dua bagian penting yaitu *dokumen training* sebanyak 3 dokumen dan *dokumen testing (dokumen uji)* yang berjumlah 1 dokumen Perguruan Tinggi. Dimana label atau kategori yang digunakan terdapat 3 yaitu ***Semarang Tengah***, ***Semarang Timur*** dan ***Semarang Selatan***. Dengan sampel query: ***Universitas Gunung Pati***.
b. Preprocessing
   Dokumen training dan uji dilakukan tokenizing, yaitu pemecahan dokumen menjadi frase atau term (kata), sesuai dengan dokumen masing – masing. Setelah dilakukan tokenizing, tahapan selanjutnya adalah di lakukan Stopword Removal atau penghilangan kata yang dirasa tidak diperlukan. Kemudian tahapan yang terakhir adalah di Stemming, yaitu penghilangan imbuhan sehingga menjadi kata dasar.

Tabel 1. Hasil Tokenization Dokumen

TERM	
universitas	swasta
dian	terbaik
nuswantoro	indonesia
udinus	diponegoro
sebuah	undip
perguruan	negeri
tinggi	tembalang
lokasi	timur
semarang	1994
tengah	unnes
diri	selatan
tahun	ngaliyan
1996	gunung pati
akreditasi	b
a	1995
merupakan	

   Pada tabel 1 menunjukkan hasil tokenization 4 dokumen dari 3 dokumen training dan 1 dokumen query, yang sudah dijadikan tiap frase dalam tiap dokumen tersebut.
c. Menentukan bobot untuk setiap term dari 4 dokumen yang terlibat dengan cara menghitung TF – IDF (Term Frequency – Inverse Document Frequency) terlebih dahulu. Pada tabel 2, pembobotannya dihitung dengan cara nilai TF dikalikan dengan nilai IDF.

Tabel 2. Tabel Perhitungan TF-IDF

d. Menghitung kemiripan vektor dokumen query dengan setiap dokumen yang sudah terklasifikasi (D1, D2, D3). Kemiripan antar dokumen menggunakan rumus Cosine Similarity sebagai berikut:

TERM	TF				DF	IDF	Wdt= TF*IDF			
	D1	D2	D3	Query			D1	D2	D3	Query
universitas	2	2	1	1	4	0	0	0	0	0
dian	1	0	0	0	1	0,60206	0,60206	0	0	0
nuswantoro	1	0	0	0	1	0,60206	0,60206	0	0	0
udinus	1	0	0	0	1	0,60206	0,60206	0	0	0
sebuah	1	0	0	0	1	0,60206	0,60206	0	0	0
perguruan	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
tinggi	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
lokasi	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
semarang	1	1	2	0	3	0,124939	0,124939	0,124939	0,249877	0
tengah	1	0	0	0	1	0,60206	0,60206	0	0	0
diri	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
tahun	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
1996	1	0	0	0	1	0,60206	0,60206	0	0	0
akreditasi	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
a	1	1	0	0	2	0,30103	0,30103	0,30103	0	0
merupakan	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
swasta	1	0	0	0	1	0,60206	0,60206	0	0	0
terbaik	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
indonesia	1	1	1	0	3	0,124939	0,124939	0,124939	0,124939	0
diponegoro	0	1	0	0	1	0,60206	0	0,60206	0	0
undip	0	1	0	0	1	0,60206	0	0,60206	0	0
negeri	0	1	2	0	2	0,30103	0	0,30103	0,60206	0
tembalang	0	1	0	0	1	0,60206	0	0,60206	0	0
timur	0	1	0	0	1	0,60206	0	0,60206	0	0
1994	0	1	0	0	1	0,60206	0	0,60206	0	0
unnes	0	0	1	0	1	0,60206	0	0	0,60206	0
selatan	0	0	1	0	1	0,60206	0	0	0,60206	0
ngaliyan	0	0	1	0	1	0,60206	0	0	0,60206	0
gunung pati	0	0	1	1	2	0,30103	0	0	0,30103	0,30103
b	0	0	1	0	1	0,60206	0	0	0,60206	0
1995	0	0	1	0	1	0,60206	0	0	0,60206	0

$$\cos(\Theta_{ij}) = \frac{\sum_k (d_{ik} d_{jk})}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}} \quad (1)$$

Hasil perhitungan skalar antara query dengan ketiga dokumen yang telah terklasifikasi. Hasil perkalian dari setiap dokumen dengan query dijumlahkan.

Tabel 3. Hasil Perkalian Skalar Query dengan (D1,D2,D3)

Wdt= TF*IDF			
D1	D2	D3	Query
0	0	0	0
0,60206	0	0	0
0,60206	0	0	0
0,60206	0	0	0
0,60206	0	0	0
0,124939	0,124939	0,124939	0
0,124939	0,124939	0,124939	0
0,124939	0,124939	0,124939	0
0,124939	0,124939	0,249877	0
0,60206	0	0	0
0,124939	0,124939	0,124939	0
0,124939	0,124939	0,124939	0
0,60206	0	0	0

Wq*di		
D1	D2	D3
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0

0,124939	0,124939	0,124939	0
0,30103	0,30103	0	0
0,124939	0,124939	0,124939	0
0,60206	0	0	0
0,124939	0,124939	0,124939	0
0,124939	0,124939	0,124939	0
0	0,60206	0	0
0	0,60206	0	0
0	0,30103	0,60206	0
0	0,60206	0	0
0	0,60206	0	0
0	0,60206	0	0
0	0	0,60206	0
0	0	0,60206	0
0	0	0,60206	0
0	0	0,30103	0,30103
0	0	0,60206	0
0	0	0,60206	0

	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0,090619
	0	0	0
	0	0	0
Total	**0**	**0**	**0,090619**

e. Menghitung panjang setiap dokumen, termasuk query dengan cara mengkuadratkan bobot setiap term dalam tiap dokumen, kemudian jumlahkan nilai kuadrat tersebut dan diakar.

Tabel 4. Menghitung Panjang Vektor Setiap Dokumen

	Panjang Vektor			
	D1	D2	D3	Query
	0	0	0	0
	0,362476	0	0	0
	0,362476	0	0	0
	0,362476	0	0	0
	0,362476	0	0	0
	0,01561	0,01561	0,01561	0
	0,01561	0,01561	0,01561	0
	0,01561	0,01561	0,01561	0
	0,01561	0,01561	0,062439	0
	0,362476	0	0	0
	0,01561	0,01561	0,01561	0
	0,01561	0,01561	0,01561	0
	0,362476	0	0	0
	0,01561	0,01561	0,01561	0
	0,090619	0,090619	0	0
	0,01561	0,01561	0,01561	0
	0,362476	0	0	0
	0,01561	0,01561	0,01561	0
	0,01561	0,01561	0,01561	0
	0	0,362476	0	0
	0	0,362476	0	0
	0	0,090619	0,362476	0
	0	0,362476	0	0
	0	0,362476	0	0
	0	0,362476	0	0
	0	0	0,362476	0
	0	0	0,362476	0
	0	0	0,362476	0
	0	0	0,090619	0,090619
	0	0	0,362476	0
	0	0	0,362476	0
Total	**2,78405**	**2,149716**	**2,468402**	**0,090619**
Akar	**1,668547**	**1,466191**	**1,571115**	**0,30103**

f. Menerapkan rumus cosine similarity, untuk menghitung kemiripan query dengan D1, D2, dan D3.

Tabel 5. Menghitung kemiripan dokumen uji dengan dokumen training

Cos (Query, Di)	Nilai	Urutan	Kelas
Cos (Query, D1)	0	2	Semarang Tengah

Cos(Query, D2)	0	3	Semarang Timur
Cos (Query, D3)	0,191602774	1	Semarang Selatan

g. Mengurutkan hasil perhitungan kemiripan

Tabel 6. Urutan hasil perhitungan

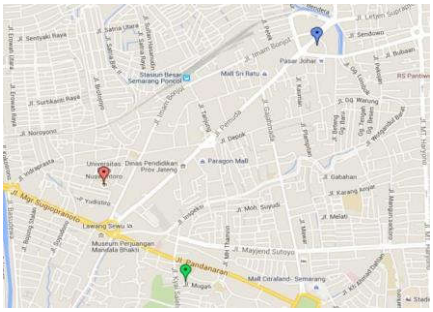
Dok	D1	D2	D3
Ranking	2	3	1

h. Dari tabel diatas ambil sebanyak k (k=1) yang paling tinggi tingkat kemiripannya dengan query dan menentukan kelas dari query, hasilnya: ***D3***. Sehingga dapat dilihat kembali bahwa ***D3*** merupakan kategori ***Semarang Selatan***. Oleh karena itu dapat disimpulkan bahwa query termasuk kedalam kategori ***Semarang Selatan***.

i. Tampilan dari desain input aplikasi pencarian perguruan tinggi

Gambar 3. Desain Input Aplikasi Pencarian Perguruan Tinggi

j. Layout Lokasi Map Hasil Pencarian Perguruan Tinggi



Gambar 4. Layout Lokasi Map Perguruan Tinggi

5. KESIMPULAN

Dengan adanya aplikasi Geographic Information Systems (GIS) berbasis Information Retrieval (IR) ini, dapat meningkatkan pencarian informasi lokasi perguruan tinggi, sekaligus dapat mempermudah calon mahasiswa baru untuk mengetahui letak geografis secara pasti beberapa perguruan tinggi di Semarang yang hendak dituju.

Penerapan algoritma K-Nearest Neighbor yang tepat dalam hal pencarian query yang sesuai di dalam dokumen corpus dan menggunakan API Google Maps yang akan menampilkan peta hasil query berdasarkan kemiripan (similarity) antara query tersebut dengan kategori teks.

DAFTAR PUSTAKA

[1] Hidayat, W. (2014, November 24). Pengguna Internet Indonesia Nomor Enam Dunia. Diambil kembali dari Website Kementerian Komunikasi dan Informatika Republik Indonesia (Kominfo): https://kominfo.go.id/content/detail/4286/pengguna-internet-indonesia-nomor-enam-dunia/0/sorotan_media

[2] Ristekdikti. (2016, Mei 23). Grafik Jumlah Perguruan Tinggi. Diambil kembali dari PANGKALAN DATA PENDIDIKAN TINGGI KEMENTERIAN RISET, TEKNOLOGI DAN PENDIDIKAN TINGGI: http://forlap.ristekdikti.go.id/perguruantinggi/homegraphpt

[3] Solichin, A., & Hasibuan, Z. a. (2012). Pemodelan Arsitektur Teknologi Informasi Berbasis Cloud Computing Untuk Institusi Perguruan Tinggi di Indonesia. Seminar Nasional Teknologi Informasi & Komunikasi Terapan (Semantik), 2012(20), 10–16.

[4] Aggarwal, C., & Zhai, C. (2012). A Survey of Text Classification Algorithms. Mining Text Data, 163–222. http://doi.org/10.1007/978-1-4614-3223-4_6

[5] Kanoje, S., Mukhopadhyay, D., & Girase, S. (2016). User Profiling for University Recommender System Using Automatic Information Retrieval. Procedia Computer Science, 78(December 2015), 5–12. http://doi.org/10.1016/j.procs.2016.02.002

[6] Pérez-Rodríguez, R., Anido-Rifón, L., Gómez-Carballa, M., & Mouriño-García, M. (2016). Architecture of a concept-based information retrieval system for educational resources. Science of Computer Programming, 1, 99–104. http://doi.org/10.1016/j.scico.2016.05.005

[7] Bordogna, G., Ghisalberti, G., & Psaila, G. (2012). Geographic information retrieval: Modeling uncertainty of user's context. Fuzzy Sets and Systems, 196, 105–124. http://doi.org/10.1016/j.fss.2011.04.005
[8] Dewandaru, A., Supriana, S. I., & Akbar, S. (2015). Evaluation on geospatial information extraction and retrieval: Mining thematic maps from web source. In 2015 3rd International Conference on Information and Communication Technology (ICoICT) (pp. 283–288). IEEE. http://doi.org/10.1109/ICoICT.2015.7231437
[9] Perea-Ortega, J. M., Lloret, E., Alfonso Ureña-López, L., & Palomar, M. (2013). Application of Text Summarization techniques to the Geographical Information Retrieval task. Expert Systems with Applications, 40(8), 2966–2974. http://doi.org/10.1016/j.eswa.2012.12.012
[10] Nesi, P., Pantaleo, G., & Tenti, M. (2016). Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering. Engineering Applications of Artificial Intelligence, 51, 202–211. http://doi.org/10.1016/j.engappai.2016.01.011
[11] Kauer, A. U., & Moreira, V. P. (2016). Using information retrieval for sentiment polarity prediction. Expert Systems with Applications, 61, 282–289. http://doi.org/10.1016/j.eswa.2016.05.038
[12] Fernández-Caramés, C., Serrano, F. J., Moreno, V., Curto, B., Rodríguez-Aragón, J. F., & Alves, R. (2016). A real-time indoor localization approach integrated with a Geographic Information System (GIS). Robotics and Autonomous Systems, 75, 475–489. http://doi.org/10.1016/j.robot.2015.08.005
[13] Sun, Y., & Li, S. (2016). Real-time collaborative GIS: A technological review. ISPRS Journal of Photogrammetry and Remote Sensing, 115, 143–152. http://doi.org/10.1016/j.isprsjprs.2015.09.011
[14] García Adeva, J. J., Pikatza Atxa, J. M., Ubeda Carrillo, M., & Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. Expert Systems with Applications, 41(4), 1498–1508. http://doi.org/10.1016/j.eswa.2013.08.047