Paper ini dipublikasikan di Proc. of Konferensi Nasional Sistem & Informatika 2008, pp.44-49, Inna Sindhu Beach Hotel, Sanur-Bali, Indonesia dan mendapat penghargaan sebagai Best Paper Award.

DESAIN SISTEM ANALISA SPATIO-TEMPORAL PENYEBARAN PENYAKIT TROPIS MEMAKAI WEB MINING

Bowo Prasetyo, M. Teduh Uliniansyah, Vitria Pragesjvara, Made Gunawan,
Gunarso, Ratih Irbandini, Anto Satriyo Nugroho, Dwi Handoko
Pusat Teknologi Informasi & Komunikasi, Badan Pengkajian & Penerapan Teknologi
BPPT Gedung II Lt.21 Jl.MH Thamrin No.8 Jakarta 10340
{praz, teduh, vitri, madegunawan, gunarso, irbandini, asnugorho, dwih}@inn.bppt.go.id

Abstract: *Malaria, Dengue Hemorrhagic Fever, and Bird Flu are three examples of various infectious diseases that spread in Indonesia in recent years. For example, Dengue Hemorrhagic Fever (DHF) is the main problem in tropical countries that emerges at the same time with rainy season. Especially in Indonesia, DHF is reported experiencing a quick increase in the number of case and death figure. Other than DHF, Bird Flu is also one of the most important diseases recently, which has placed Indonesia as the hotspot in the world, with the very quickly increasing and worrying number of case. Under national economic condition that is not very good, the epidemic of these infectious diseases is a serious threat to Indonesian people´s life, where many segments of society is still living under condition that is not satisfying healthy conditions. Therefore, monitoring the spread of disease, handling the disease itself, and preventif actions to prevent their spreading are a big task for Indonesian people, especially medical researchers/practitioners, and health policy makers in Indonesia.*

Keywords/Kata kunci: web mining, tropical disease, spatio-temporal analysis, visualization, GIS

1. Latar Belakang

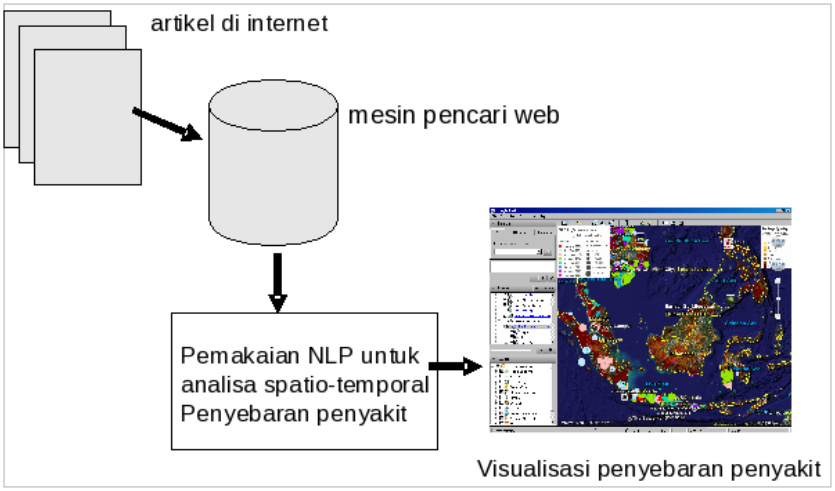
Malaria, Demam Berdarah Dengue, Flu Burung, adalah contoh dari berbagai penyakit menular yang akhir-akhir ini mewabah di Indonesia. Demam Berdarah Dengue (DBD) misalnya, merupakan masalah utama di negara tropis yang datang bersamaan dengan musim hujan. Di Indonesia khususnya, DBD dilaporkan mengalami peningkatan jumlah kasus dan angka kematian secara cepat (Utama A. 2004, 2005). Di antara berbagai penyakit yang lain, Flu Burung khususnya, menurut laporan yang dimuat di jurnal Nature mengindikasikan bahwa Indonesia merupakan salah satu hotspot dunia, dengan tingkat penyebaran yang relatif sangat pesat dan mengkhawatirkan. (Nature News 2006, Butler D. 2006a, Butler D. 2006b).

Berbagai upaya telah dilakukan untuk menganalisa mekanisme timbulnya penyakit, pola penyebaran dan penanggulangannya (Marzuki S. Verhoef J. Snippe H. 2003). Pemantauan penyebaran penyakit, terutama yang tingkat penyebarannya sangat tinggi, sangat dibutuhkan oleh peneliti, praktisi dan pengambil kebijakan di bidang kesehatan, agar dapat membuat keputusan akurat secepat mungkin. Dalam upaya tersebut, tersedianya sistem informasi spatio-temporal yang mampu memantau penyebaran penyakit di lokasi geografis tertentu pada suatu kurun waktu, merupakan kebutuhan vital. Sistem tersebut mampu memberikan informasi yang jelas, di daerah mana suatu penyakit menyebar dan seberapa jauh tingkat penyebarannya, sehingga pengambil kebijakan akan mampu memprediksi pola penyebaran dari penyakit tersebut dan sedini mungkin mengidentifikasikan daerah yang rawan terjangkiti oleh penyakit menular itu. Tetapi hingga saat ini, data yang lengkap dan informatif mengenai penyebaran penyakit di Indonesia tidak tersedia, sehingga pengembangan sistem ini tidak mudah dilakukan.

Di sisi lain, internet merupakan salah satu sumber informasi yang potensial dimanfaatkan sebagai pemasok data. Sebagai contoh, jika kata kunci “demam berdarah” dimasukkan sebagai entry pada situs http://google.co.id, akan diperoleh lebih dari 300 ribu situs yang mengandung kata-kata tersebut. Tentunya tidak semua situs mengandung informasi yang relevan. Informasi yang diberikan oleh ribuan situs tersebut harus terlebih dahulu diseleksi, dan selanjutnya dilakukan ekstraksi informasi secara terbatas, terhadap situs yang dianggap memiliki informasi yang relevan. Dalam hal ini, teknologi web mining memungkinkan proses ekstraksi informasi berjalan secara otomatis dan efektif.

Web mining merupakan salah satu teknologi di bidang komputasi yang dewasa ini berkembang dengan pesat. Web mining bertujuan mengekstrak informasi yang sangat berharga dari data yang disajikan di internet dan berskala besar. Teknologi web mining sering dipakai pada situs online shopping, misalnya Amazon.com, untuk menganalisa perilaku pengunjung situs dan pembeli produk. Informasi hasil olahan web mining ini akan dipakai sebagai masukan dalam perancangan strategi marketing, untuk meningkatkan profit dari perusahaan tersebut (Chakrabarti S. 2003). Dalam studi ini, web mining dipakai untuk menemukan informasi spatio-temporal penyebaran penyakit dari informasi yang tersedia di internet. Informasi spatio temporal ini berfungsi sebagai data input bagi pengembangan sistem pemantauan penyebaran penyakit tropis di Indonesia.

2. Rancangan sistem



Gambar 1. Rancangan sistem analisa spatio-temporal penyebaran penyakit tropis

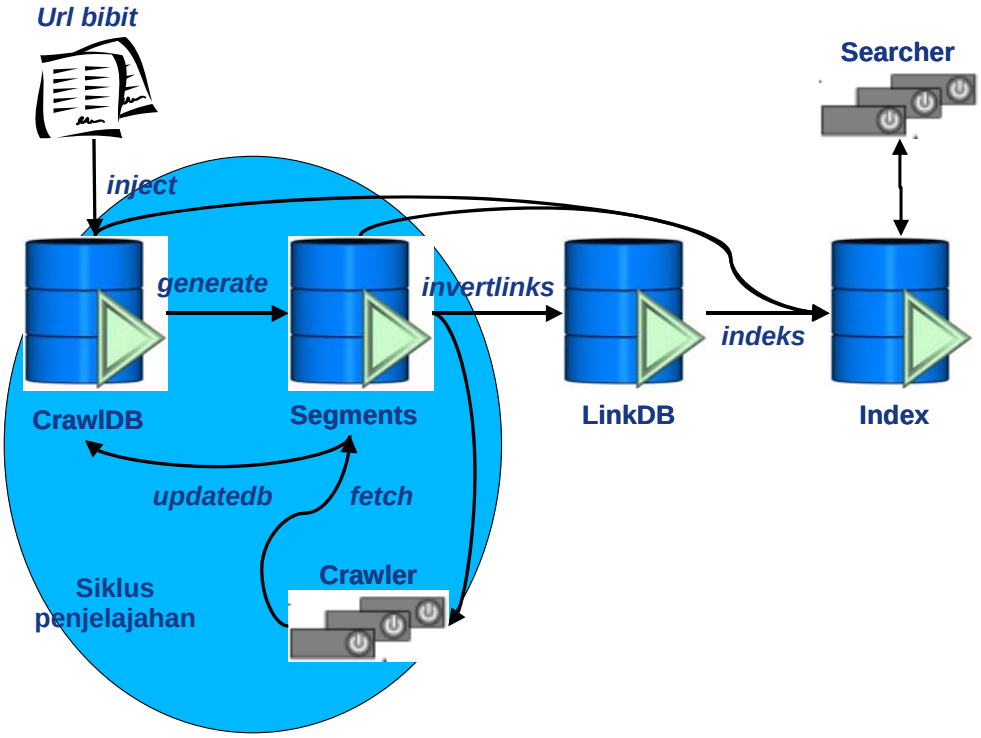
Gambar 1 memperlihatkan skema sistem yang akan dibangun dalam penelitian ini, yang terdiri dari: mesin pencari web untuk mengunduh data dari internet, menyimpannya dalam database lokal, lalu melakukan pencarian secara offline; text mining dimana teknologi Natural Language Processing (NLP) akan dipakai untuk mengekstrak informasi spatio-temporal penyebaran penyakit dan menganalisanya; dan visualisasi penyebaran penyakit.

2.1 Pembuatan database tekstual memakai Nutch

Bagian pertama dari sistem web mining adalah sistem mesin pencari web yang memanfaatkan perangkat lunak open source *Nutch* (Mike Cafarella and Doug Cutting 2004) (Url pengunduhan Nutch). Fungsi utama dari sistem Nutch ini ada dua, yang pertama adalah fungsi *crawling*, yaitu mengumpulkan data tekstual dari internet lalu menyimpannya di database; dan yang kedua adalah fungsi *searching*, yaitu mencari informasi yang diperlukan dari dalam database tersebut. Dengan crawling data-data secara umum dalam jumlah besar diunduh dari internet dan disimpan di database lokal sehingga dapat dimanfaatkan secara offline. Sedangkan fungsi searching memungkinkan pencarian dari dalam tumpukan data-data tersebut, informasi tertentu yang diperlukan oleh penelitian ini untuk diteruskan ke bagian selanjutnya untuk dianalisa.

Nutch adalah perangkat lunak mesin pencari web open source berbasis Java yang dibangun di atas Lucene Java, yaitu sebuah mesin pencari teks open source yang berfitur lengkap. Sebagai mesin pencari web, Nutch menambahkan di atas Lucene Java beberapa fungsi yang spesifik web, seperti *crawler*, database *link-graph*, serta *parser* untuk HTML dan format dokumen lainnya seperti JavaScript, MS Excel-PowerPoint-Word, Open Office, Adobe PDF, Shockwave Flash, RSS, ZIP dll.

Sebagaimana ditunjukkan dalam Gambar 2, mula-mula seperangkat url bibit disuntikkan dengan proses *inject* ke dalam database *CrawlDB* (disebut juga *WebDB*), yang berfungsi untuk menyimpan seluruh url hasil crawling beserta informasi pelengkapnya. Proses inject hanya dilakukan satu kali di awal saja, dan url bibit ini akan menjadi titik awal bagi crawler untuk memulai penjelajahan dan pengunduhan web. Selanjutnya dilakukan proses *generate* untuk menghasilkan daftar url dan menyimpannya ke dalam database *Segment* yang akan dijelajahi oleh crawler. Setelah itu proses *fetch* akan dijalankan untuk menjelajahi dan mengunduh url di dalam daftar. Hasil pengunduhan berupa konten setiap url akan disimpan di dalam Segment yang diberi nama sesuai dengan tanggal dan waktu mulainya penjelajahan. Berikutnya dilakukan proses *updatedb* untuk memperbarui CrawlDB dengan url-url yang baru diunduh di dalam Segment, sehingga url di dalam CrawlDB akan bertambah dan berisi gabungan url sebelumnya dengan url-url yang baru. Selanjutnya proses akan berulang kembali ke generate – fetch – updatedb – generate – ... demikian seterusnya menjadi sebuah siklus penjelajahan yang berkesinambungan.



Gambar 2. Arsitektur mesin pencari web Nutch

Setelah jumlah url di dalam CrawlDB dianggap cukup maka siklus penjelajahan di atas dihentikan, lalu dilanjutkan dengan proses pengolahan data supaya siap dilakukan pencarian. Dimulai dari proses *invertlinks* untuk menganalisa struktur link di antara setiap url di dalam Segment dan menyimpan hasilnya di dalam database *LinkDB*. Kemudian dilanjutkan dengan proses *index* untuk mengindeks kata-kata di dalam Segment dan menghubungkannya dengan setiap url yang tersimpan di CrawlDB dan LinkDB, lalu menyimpan hasilnya di dalam database *Indexes*. Berikutnya perlu dilakukan proses *dedup* pada Indexes untuk menghapus indeks ganda yang mungkin terbuat pada proses index sebelumnya. Dan terakhir adalah proses *merge* untuk menggabungkan beberapa Indexes menjadi satu dan menyimpannya di dalam database *Index*. Selanjutnya maka terhadap data yang tersimpan di dalam empat jenis database tersebut, yaitu Index, LinkDB, Segment dan CrawlDB telah siap dilakukan pencarian.

Ada beberapa cara untuk melakukan pencarian terhadap database Nutch. Yang pertama dan paling umum dilakukan adalah melalui antarmuka web berbasis *JSP* yang telah tersedia di dalam paket perangkat lunak Nutch. Antarmuka web ini dapat dijalankan di web server yang mensupport Java servlet container, seperti open source *Apache Tomcat* yang turut dipaket menjadi satu dengan Nutch. Cara kedua adalah melalui perintah command line Java yang telah disediakan, misalnya `nutch org.apache.nutch.searcher.NutchBean keyword`. Akhirnya kita juga dapat melakukan pencarian terhadap database Nutch secara programatikal dalam bahasa Java dengan memanfaatkan *Nutch API* yang tersedia, seperti CrawlDbReader, SegmentReader, LinkDbReader, IndexReader, Searcher, NutchBean dll.

Di dalam penelitian ini database Nutch diakses secara programatikal menggunakan perangkat lunak *GwiNutch*, sebuah program GUI berbasis Java yang sedang dikembangkan oleh sebagian tim kami untuk mengambil data dari database Nutch, kemudian secara default menggunakan open source *Weka* (atau secara opsional program lainnya, misalnya text mining dengan NLP seperti di bawah) untuk melakukan berbagai proses *data mining* terhadap data tersebut, kemudian menampilkannya kepada pengguna.

2.2 Ekstraksi informasi spatio-temporal penyebaran penyakit

Bagian kedua dari sistem web mining adalah pemakaian text mining untuk mengekstrak informasi spatio-temporal suatu penyakit dari informasi berupa teks yang berasal dari web. Spatio-temporal maksudnya dimana penyakit itu ditemukan dan kapan terjadi, berapa banyaknya korban, dan berapa yang meninggal. Sistem ekstraksi informasi yang akan dikembangkan pada sistem yang sedang dibangun terdiri dari beberapa komponen, yaitu (Jackson P., Moulinier I. 2002)

- *Tokenizer*

  Tokenizer adalah sebuah modul yang digunakan untuk mendapatkan kata berdasarkan delimiter yang digunakan dalam suatu bahasa alami. Modul ini penting bagi bahasa-bahasa alami yang tidak menggunakan spasi sebagai pemisah antara satu kata dengan kata lainnya dalam suatu kalimat seperti bahasa Jepang, China, Thailand, dll. Dalam

hal bahasa Indonesia dimana spasi digunakan untuk memisahkan kata-kata dalam suatu kalimat, modul ini bisa dimasukkan dalam modul *POS Tagger* (pemarka jenis kata).

- *POS Tagger*

POS Tagger atau pemarka jenis kata adalah modul yang bertugas untuk menentukan jenis kata apa yang dimiliki oleh sebuah kata. Sebagaimana diketahui, morfologi kata-kata dalam bahasa Indonesia adalah suatu proses yang sangat kompleks dengan menggunakan kombinasi beragam imbuhan awalan, sisipan, akhiran, maupun konfiks (gabungan imbuhan). Kompleksnya proses pembentukan kata dalam bahasa Indonesia ditambah lagi dengan banyaknya imbuhan-imbuhan asing yang sering digunakan seperti supra-, intra-, inter-, infra-, dll. Belum lagi jika ditambahkan dengan pengaruh bahasa daerah seperti bahasa Jawa atau Betawi yang terkadang digunakan dalam bahasa tulisan. Membuat suatu kamus elektronik yang memuat semua kata (berimbuhan) akan membutuhkan sumber daya waktu dan tenaga yang tidak sedikit. Adanya modul pemarka jenis kata akan sangat membantu dalam pembuatan kamus elektronik dan sangat berguna bagi modul analisa kalimat dalam menentukan bagian-bagian kalimat yang berfungsi sebagai subyek, predikat, obyek, dan keterangan kalimat.

- *Regex Matcher*

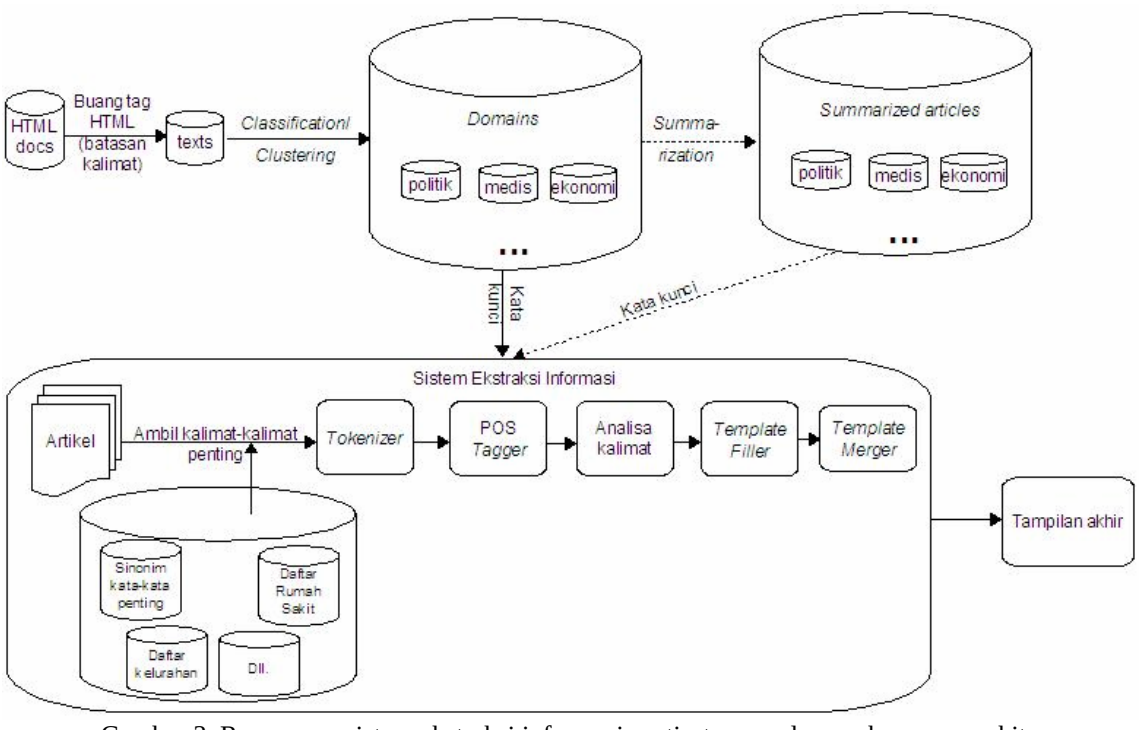
Regex Matcher adalah kependekan dari *regular expression matcher*, yaitu sebuah modul yang bertugas mencari pola-pola tertentu yang ingin dicari dalam suatu kata atau kalimat.

- *Template Filler*

Template Filler adalah suatu modul yang mempunyai tugas mencari informasi yang dirasakan berharga dalam kalimat/artikel untuk kemudian dimasukkan dalam suatu tabel *template*.

- *Template Merger*

Template Merger adalah suatu modul yang bertugas menggabungkan beberapa tabel *template* yang mempunyai kemiripan untuk mendapatkan gambaran informasi yang lebih lengkap.



Gambar 3. Rancangan sistem ekstraksi informasi spatio-temporal penyebaran penyakit

Sebagaimana tertera pada Gambar 3, berikut adalah keterangan ringkas mengenai sistem ekstraksi informasi yang akan dikembangkan. Sistem ekstraksi informasi memiliki input berupa kumpulan artikel-artikel teks yang diterima dari modul klasifikasi/*clustering* dan peringkas dokumen (*summarization module*). Kalimat-kalimat dalam artikel-artikel tersebut kemudian akan dipilih lebih lanjut dengan mencari kalimat-kalimat yang mempunyai kata-kata penting yang dibutuhkan untuk mengisi tabel *template* dengan merujuk sekumpulan *data base* sinonim kata (yang dirasa penting), daftar kelurahan, kecamatan, kabupaten, dan rumah sakit se-Indonesia.

Kalimat-kalimat yang terpilih untuk diproses kemudian diproses oleh modul pemarka jenis kata untuk menentukan jenis kata. Sebagaimana telah disebutkan sebelumnya, tugas modul *tokenizer* akan dijalankan oleh modul pemarka jenis kata

karena kalimat-kalimat dalam Bahasa Indonesia menggunakan spasi sebagai pemisah kata, Daftar kata berikut informasi mengenai jenis kata akan dimasukkan dalam kamus elektronik modul analisa kalimat.

Proses selanjutnya adalah menentukan subyek, predikat, obyek, dan keterangan kalimat yang akan dilakukan oleh modul analisa kalimat. Modul analisa kalimat akan menggunakan perangkat lunak Link Grammar yang dikembangkan oleh Sleator dan Temperley. Link Grammar adalah perangkat lunak penganalisa sintaksis bahasa Inggris berbasis aturan (*rule-based parser*). Saat ini, Link Grammar dikembangkan oleh Abi Word sebagai alat bantu pemeriksa kebenaran tata bahasa Inggris. Link Grammar sudah dikembangkan sejak tahun 1993 dan saat ini mencapai versi 4.3.5. Aturan-aturan (rules) yang ada pada Link Grammar akan dicoba dirubah dan dikembangkan untuk menganalisa kalimat-kalimat yang ditulis dalam Bahasa Indonesia. Hal ini merupakan suatu tugas yang tidak mudah mengingat kompleksnya aturan-aturan (*rules*) yang ada dalam Link Grammar karena sudah dikembangkan sejak tahun 1993.

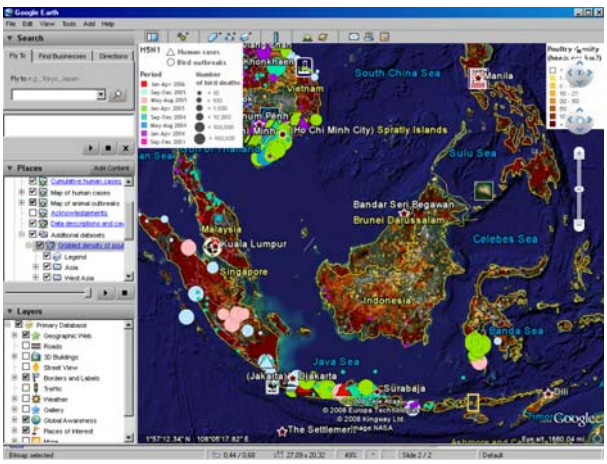
Hasil keluaran modul analisa kalimat akan diterima oleh modul *Template Filler* yang akan menentukan bagian-bagian kalimat yang dirasakan perlu dimasukkan dalam tabel *template*. Selanjutnya modul Template Merger akan menggabungkan tabel-tabel *template* yang data-datanya mempunyai kemiripan. Hasil akhir dari sistem ekstraksi informasi adalah tabel-tabel berisikan data-data mengenai suatu penyakit (tanggal, jumlah penderita, lokasi, jenis penyakit, dll.) yang kemudian akan digunakan oleh sebuah modul visualisasi.

2.3 Visualisasi memakai Google Earth

Bagian akhir dari sistem web mining adalah visualisasi spatio-temporal penyebaran penyakit menular. Salah satu alternatif yang sedang dikaji adalah pemanfaatan Google Earth, yaitu sebuah program virtual yang disebut Earth Viewer seperti ditunjukkan pada Gambar 4. Program ini memetakan bumi dari gambar yang dikumpulkan dari pemetaan satelit, fotografi udara dan GIS 3D. Google Earth mendukung pengelolaan data Geospatial tiga dimensi. Data tersebut ditampilkan melalui Keyhole Markup Language (KML) yaitu bahasa berbasis XML untuk menampilkan anotasi geografis dan visualisasi peta 2D dan browser bumi 3D, membuat model dan menyimpan fitur geografis seperti titik, garis, gambar, poligon serta model untuk ditampilkan di Google Earth, Google Maps dan lain lain.

Sebuah file yang ditulis dalam KML menspesifikasikan satu set fitur yang terdiri dari tanda tempat, citra, polygon, model 3D, deskripsi teks. Format KML diproses oleh Google Earth dengan cara yang sama seperti pemrosesan HTML dan XML. Dan seperti HTML, KML juga memiliki struktur berbasis tag dengan nama dan atribut yang digunakan untuk tujuan tampilan tertentu. Contoh file KML adalah seperti di bawah:

```
<Create>
<Update>
<targetHref>http://myserver.com/Point.kml</targetHref>
<Create>
<Document targetId="region24">
<Placemark id="placemark891">
<Point>
<coordinates>-95.48,40.43,0</coordinates>
</Point>
</Placemark>
</Document>
</Create>
</Update>
```



Gambar 4. Tampilan pada Google Earth yang memperlihatkan data penyebaran penyakit

3. Kesimpulan dan Penelitian Selanjutnya

Sistem web mining yang kami kembangkan memungkinkan analisa dan visualisasi penyebaran penyakit menular yang tercatat di situs-situs di internet secara otomatis dan efisien. Nutch di bagian hulu yang bertugas sebagai pengumpul data dari internet, akan memasok data tersebut kepada program text mining dengan NLP yang akan menganalisa pola penyebaran penyakit, dan akhirnya menyerahkan hasilnya kepada Google Earth di hilir untuk ditampilkan secara spatio-temporal.

Tersedianya sistem informasi spatio-temporal seperti ini yang mampu memantau penyebaran penyakit di lokasi geografis tertentu pada suatu kurun waktu adalah kebutuhan vital. Dengan sistem yang mampu memberikan informasi yang jelas, di daerah mana suatu penyakit menyebar dan seberapa jauh tingkat penyebarannya, maka para pengambil kebijakan akan mampu memprediksi pola penyebaran dari penyakit tersebut dan sedini mungkin mengidentifikasikan daerah yang rawan terjangkiti oleh penyakit menular itu.

Di masa depan, penelitian ini akan melakukan pemodelan matematika untuk penyebaran penyakit menular, melanjutkan pengumpulan data dari internet sebanyak mungkin, mengimplementasikan program text mining dengan NLP, dan menelaah lebih jauh proses visualisasi data dengan Google Earth.

4. Daftar Pustaka

[1]Utama, A (2004). Dengue: Permasalahan dan Solusinya, *Harian Kompas*, 27 Februari 2004 (http://www.kompas.com/kompas-cetak/0402/27/humaniora/880208.htm)
[2]Utama, A (2005). Menantikan Vaksin DBD, *Republika*, 19 Februari 2005 (http://www.republika.co.id/kolom_detail.asp?id=188167&kat_id=16)
[3]Nature News (2006). Indonesia is currently the world's avian-flu hotspot, *Nature* 440, 726-727, April 2006
[4]Butler D. (2006a). Indonesian bird-flu cluster rings alarm bells, *Nature* 441, 554-555, 1 June 2006
[5]Butler D. (2006b). Bird flu outbreaks in Indonesia going unstudied, *Nature*, 28 July 2006
[6]Marzuki S, Verhoef J, Snippe H. Eds. (2003). Tropical Diseases From Molecule to Bedside, *Advances in Experimental Medicine and Biology*, Vol.531
[7]Chakrabarti S. (2003). Mining the Web Discovering Knowledge from Hypertext Data, Morgan Kaufmann-Elsevier Science
[8]Mike Cafarella and Doug Cutting (2004). Building Nutch: Open Source Search, *ACM Queue*, vol. 2, no. 2, April 2004
[9]Url pengunduhan Nutch: http://lucene.apache.org/nutch/
[10]Jackson P., Moulinier I. (2002). Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization (Natural Language Processing, 5), John Benjamins Publishing Co