Paradikma Text Mining

Nur Rosyid M

Mesin Pencari

Google  **Lebih manusiawi**

Menganalisa pencarian

Definisi

Text Mining adalah proses **penemuan akan informasi** atau **trend baru** yang **sebelumnya tidak terungkap** dengan **memproses dan menganalisa data** dalam *jumlah besar*.

Text mining, also known as intelligent text analysis, text data mining , unstructured data management, or knowledge discovery in text …, refers generally to the process of extracting interesting and non-trivial information and knowledge (usually converted to metadata elements) from unstructured text (i.e. free text) stored in electronic form. (*wikipedia*)

Data Mining Vs Text Mining

	Data Mining	Text Mining
Data Besar	✓	✓
Dimensi tinggi	✓	✓
Data dan Struktur berubah-ubah	✓	✓
Noise	✓	✓
Data Terstruktur	✓	✗

Definisi

Text adalah keranjang kata

Text mining mengolah data berupa **text yang complex** dan **tidak lengkap**, **arti yang tidak jelas** dan **tidak standard**, dan bahasa yang **berbeda** ditambah **translasi yang tidak akurat**.

Definisi

Text di ciptakan untuk dikonsumsi manusia, bukan untuk di gunakan oleh mesin, maka perlu **NLP (Natural Language Processor) dan CL (Computational Linguistics)** memproses unstructured text guna medapatkan egekstraksi informasi.

Text mining lebih dari information retrieval atau sekedar mesin pencari . Tetapi text mining menggunakan information retrieval untuk menyaring dan mengurangi jumlah informasi untuk diproses selanjutnya.

Definisi

Text Mining adalah proses penemuan kembali relasi dan fakta yang terkubur didalam text, bukan hanya **klasifikasi Text** atau **pencarian kata-kata**

Aplikasi text mining

Pada ustructured text dalam bentuk emails, intant messages, dan blogs, pada umumnya penguna ingin mencari atau “mine” informasi mengenai orang (seperti email pengirim, alamat, nama lengkap, dll), perusahaan (seperti nama lengkap dan lokasi), organisasi, dan kejadian-kejadian (seperti penemuan baru, pengumuman penting, dll)

Pada berita dari berbagai sumber, text mining bisa di gunakan untuk membandingkan berita yang sama atau berbeda yang berasal dari sumber yang berbeda, mungkin dengan bahasa yang berbeda, mengalisa berita berdasarkan waktu ,menganalisa trend riset.

Aplikasi text mining

Untuk technical working paper, documentasi, dan software specifikasi dokumen, text mining bisa di gunakan untuk mengekstrak software requirement dari spesifikasi dokumen secara otomatis atau mendeteksi ke kurangan antarasource code dan documentasinya secara otomatis.

Untuk web pages, text mining bisa di gunakan untuk menganalisa website perusahaan, structur websitenya, perbandingan website content yang satu dengan site yang lain. Masih banyak lagi aplikasi text mining yang di butuhkan.

Proses Text Mining

Proses text mining mencakup beberapa sub-task, seperti information retrieval, categorization, POS tagging, Clustering, dan lainnya, yang bisa di katergorikan kedalam framework “Knowledge Discovery in Databases” (KDD)

Proses KDD terdiri dari :

- ✓ Selection
- ✓ Preprocessing
- ✓ Transformation
- ✓ Data Mining
- ✓ Interpretation/Evaluation

Selection

Tujuan information retrieval adalah untuk mengubah **unstructured text** menjadi **structured data** atau format yang mudah untuk di proses lebih lanjut nantinya.

Contoh : email body di proses secara automatic untuk mendapatkan nama, email address, alamat, telephone, dan information yang relevan lainnya.

Text mining yang bisa masuk dalam phase ini termasuk **Information Retrieval, Categorization**, dan **Clustering**

Preprocessing

Preprocessing memfocuskan pada data cleaning, termasuk menghilangkan noise di data, atau mengadaptasi noise, dan mengatasi informasi yang hilang atau tidak Komplit

Preprocessing:

POS (Part of Speech) Tagging yang tujuannya memberikan label pada setiap kata dalam kalimat dan mengasosiasikan dengan “speech” yang relevan.

Disambiguation adalah activitas untuk menentukan arti atau sense akan kata-kata yang tidak jelas atau ambiguos

Transformation

Tranformation step bertujuan untuk menemukan *fitur-fitur yang tersimpan di dalam data yang penting* berdasarkan kebutuhan yang diperlukan, dengan mengurangi **jumlah variabel dan data** yang tidak terlalu di perlukan.

Tranformation meliputi Disambiguation dan Clustering

Clustering adalah activitas untuk menciptakan model yang bisa digunakan untuk meng-index dokumen pada tahap yang berbeda. Termasuk didalamnya thesaurus atau ontology.

Misal kata meja berhubungan dg kata apa saja.

Data Mining

Process Data Mining bertujuan untuk menghasilkan patterns yang berguna dari koleksi text. Activitas text mining untuk step data mining terdiri dari pemilihan mining teknik yang benar, penentuan mining model dan parameters.

Clustering dan Parsial Parsing bisa dimasuk dalam step ini.

Data Mining

Parsial parsing atau robust parsing bertujuan untuk *mengidentifikasi relationship yang lebih dalam antar kata-kata dalam kalimat*. Membutuhkan hasil dari POS Tagging dan biasanya di gunakan secara bersamaan.

Teknik pengunaan rule based system, memory based system, statistical method, atau kombinasi antar teknik banyak di gunakan untuk parsial parsing.

Interpretation/Evaluation

Pada tahap ini adalah text Summarization yang bertujuan mencari *key content* yang dapat bisa merepresentatikan keseluruhan text secara akurat

Text summarization di gunakan untuk menjelaskan seluruh kontent text dengan mengekstrak hanya keyword yang penting, untuk menghidari membaca seluruh text, atau untuk membantu proses text searching supaya lebih cepat dan akurat dengan memfocuskan hanya pada keyword penting

Kesimpulan

- ✓ Dengan terus meningkatnya jumlah “digitized textual media” menunjukan nyatanya tantangan akan “overload” akan informasi dan pentingnya bidang text mining.
- ✓ Kita memerlukan tak hanya text mining system, tapi juga knowledge management system di bantu dengan robust text mining software untuk *mengekstrak, memprocess, me-mine, mengorganisasi, dan memonitor textual data dalam jumlah besar.*
- ✓ Solusi text mining harus lebih dari **sekedar efektif search, akurat natural language processor, dan text summization,** text mining harus memiliki kemampuan untuk menemukan fakta dan relationship yang baru yang sulit di dapat tanpa text mining