KOM341
Temu Kembali Informasi

KULIAH #9
- Text Summarization

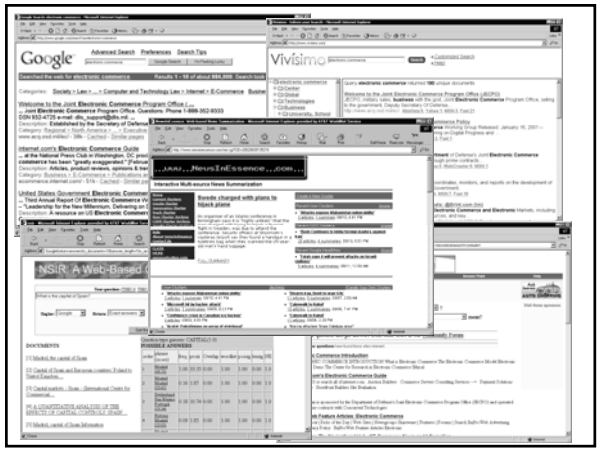
Sumber

Tutorial
ACM SIGIR
Sheffield, UK
July 25, 2004

Dragomir R. Radev
CLAIR: Computational Linguistics And Information Retrieval group
University of Michigan

Information overload

- Masalah:
  - 4 Billion URL yang di-indeks oleh Google
  - 200 TB data dalam Web [Lyman & Varian, 2003]
- Pendekatan yang mungkin:
  - information retrieval
  - document clustering
  - information extraction
  - question answering
  - text summarization



Contoh artikel

MILAN, Italy, April 18. A small airplane crashed into a government building in heart of Milan, setting the top floors on fire, Italian police reported. There were no immediate reports on casualties as rescue workers attempted to clear the area in the city's financial district. Few details of the crash were available, but news reports about it immediately set off fears that it might be a terrorist act akin to the Sept. 11 attacks in the United States. Those fears sent U.S. stocks tumbling to session lows in late morning trading.

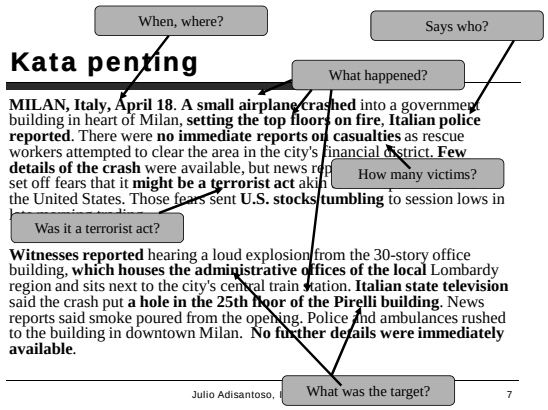
Witnesses reported hearing a loud explosion from the 30-story office building, which houses the administrative offices of the local Lombardy region and sits next to the city's central train station. Italian state television said the crash put a hole in the 25th floor of the Pirelli building. News reports said smoke poured from the opening. Police and ambulances rushed to the building in downtown Milan. No further details were immediately available.

Kata penting

MILAN, Italy, April 18. **A small airplane crashed** into a government building in heart of Milan, **setting the top floors on fire**, **Italian police reported**. There were **no immediate reports on casualties** as rescue workers attempted to clear the area in the city's financial district. **Few details of the crash** were available, but news reports about it immediately set off fears that it **might be a terrorist act** akin to the Sept. 11 attacks in the United States. Those fears sent **U.S. stocks tumbling** to session lows in late morning trading.

Witnesses reported hearing a loud explosion from the 30-story office building, **which houses the administrative offices of the local** Lombardy region and sits next to the city's central train station. **Italian state television** said the crash put **a hole in the 25th floor of the Pirelli building**. News reports said smoke poured from the opening. Police and ambulances rushed to the building in downtown Milan. **No further details were immediately available**.

Kata penting



MILAN, Italy, April 18. **A small airplane crashed** into a government building in heart of Milan, **setting the top floors on fire**. **Italian police reported**. There were **no immediate reports of casualties** as rescue workers attempted to clear the area in the city's financial district. **Few details of the crash** were available, but news rep... set off fears that it **might be a terrorist act** akin ... the United States. Those fears sent **U.S. stocks tumbling** to session lows in late morning trading.

Witnesses reported hearing a loud explosion from the 30-story office building, **which houses the administrative offices of the local** Lombardy region and sits next to the city's central train station. **Italian state television** said the crash put **a hole in the 25th floor of the Pirelli building**. News reports said smoke poured from the opening. Police and ambulances rushed to the building in downtown Milan. **No further details were immediately available**.

Jenis Ringkasan

- Tujuan
  - Indicative, informative, critical summaries
- Bentuk
  - Ekstrak (paragraf/kalimat/frase)
  - Abstrak: suatu ringkasan yg padat dari topik permasalahan dari suatu dokumen [Paice90].
- Dimensi
  - Single-document vs. multi-document
- Konteks
  - Query-specific vs. query-independent

Jenis Ringkasan

- Indikatif vs. informatif
  digunakan untuk kategorisasi secara cepat vs. pemrosesan isi.
- Ekstrak vs. abstrak
  daftar fragmen teks vs. menyimpulkan kembali isi secara koheren.
- Query-independen vs. query-spesifik
  mengikuti pandangan penulis vs merefleksikan minat dari user
- Background vs. just-the-news
  asumsikan jika pengetahuan pembaca sebelumnya tidak banyak vs. sangat mengikuti perkembangan.
- Single-dokumen vs. multi-dokumen
  berdasarkan pada satu teks vs. penggabungan beberapa teks.

Hasil Ringkasan

- headlines
- outlines
- minutes (notulen)
- biographies
- sound bites
- movie summaries
- chronologies, etc.

Ekstrak vs Abstrak

- Ringkasan Teks

  Proses penyaringan informasi yang paling penting dari suatu sumber (atau beberapa sumber) untuk menghasilkan suatu versi yang ringkas untuk user.
- Extract vs. Abstrak
  - Suatu extract adalah ringkasan yang isi seluruhnya disalin dari input.
  - Suatu abstrak adalah ringkasan yang paling sedikit ada isinya yang tidak ada pada input, mis. kategorisasi topik, menyarikan kembali isi, dsb.

Pendekatan Tradisional

- Human summarization and abstracting
- Professional abstractors

Pendekatan Statistika

- Linear Feature Combination untuk melakukan ekstraksi kalimat.

$$Weight(U) = \alpha \cdot Loc(U) + \beta \cdot Phrase(U) + \gamma \cdot Thema(U) + \delta \cdot Term(U)$$

Linear Feature Combination

- Location : bobot diberikan pada suatu unit teks berdasarkan posisi munculnya di awal, tengah, atau akhir paragraf atau seluruh dokumen, atau pada bagian tertentu dari dokumen seperti pada bab pengenalan atau kesimpulan. (mis, judul, introduksi, kesimpulan).
- Phrase: bobot diberikan pada suatu unit teks bila fixedphrases muncul, misal : (singkatnya, penyelidikan kami menunjukkan, tujuan artikel ini adalah, ...), kata penekanan (penting, terutama, ...)
- Thema : bobot diberikan pada suatu unit teks karena adanya kata yang penting secara statistik (mis. Kata tf.idf) pada unit tersebut sesuai dengan dokumennya
- Term : bobot diberikan pada suatu unit teks untuk kata-katanya yang juga muncul di judul, berita utama, paragraf awal, atau query dari user.

Contoh Sederhana

Ancaman krisis pangan sudah di depan mata dan bakal dirasakan rakyat bila pemerintah tidak segera turun tangan untuk memperbaiki sektor **pertanian** yang juga mengalami **krisis**. Tanda akan **krisis pangan** sudah tampak dan diungkapkan Jafar Hasnah dalam acara sosialisasi program perluasan areal tanam tahun 2004 di Balikpapan. Salah satu **ancaman** bagi **stabilitas pangan** nasional, antara lain, adalah semakin menyempitnya lahan **pertanian** akibat **alih fungsi** untuk **perumahan** dan tempat usaha lain, seperti **pabrik** dan **pergudangan**. Selain itu, irigasi **pertanian** di berbagai daerah semakin tidak mendapat perhatian dalam hal perawatan sehingga di sana-sini retak, bocor, dan tidak lagi efektif untuk mengalirkan air ke petak-petak areal **pertanian**.

ancaman	krisis	pangan	pertanian	stabilitas
20	25	15	100	10

alih fungsi	perumahan	pabrik	gudang
10	15	5	5

Tabel menunjukkan banyaknya dokumen yang mengandung kata. Misalkan N=1000 dokumen. Ringkas menjadi 2 kalimat dengan pembobotan TF.IDF.