RELEVANCE FEEDBACK

Oleh : Rahmat Robi Waliyansyah, M.Kom.

RELEVANCE FEEDBACK

- Relevance feedback adalah umpan balik dari user yang tujuannya mengetahui apakah data yang terretrieve menurut user sesuai atau relevan dengan query yang dimaksudkan.
- Terdapat 3 metode relevance feedback :

1. User judgement (explicit relevance feedback)
2. User behavior (implicit relevance feedback)
3. Top K retrieved (blind/pseudo relevance feedback)

USER JUDGEMENT (EXPLICIT RELEVANCE FEEDBACK)

- Explicit relevance feedback diperoleh dari penilai relevan dari relevansi suatu dokumen yang diambil dengan query tertentu.
- Penilai relevan yang dimaksud adalah penilaian yang ditafsirkan sendiri oleh user atau pengguna.
- Penilai relevan ini dapat menggunakan biner atau system relevansi bergradasi.

User behavior (implicit relevance feedback)

- Implicit relevance feedback disimpulkan melalui perilaku pengguna.
- Seperti mencatat dokumen-dokumen yang mereka lakukan dan dokumen yang tidak dipilih untuk dilihat, durasi waktu yang digunakan untuk melihat dokumen, atau browsing atau scrolling halaman.

PERBEDAAN IMPLICIT DENGAN EXPLICIT MENCAKUP

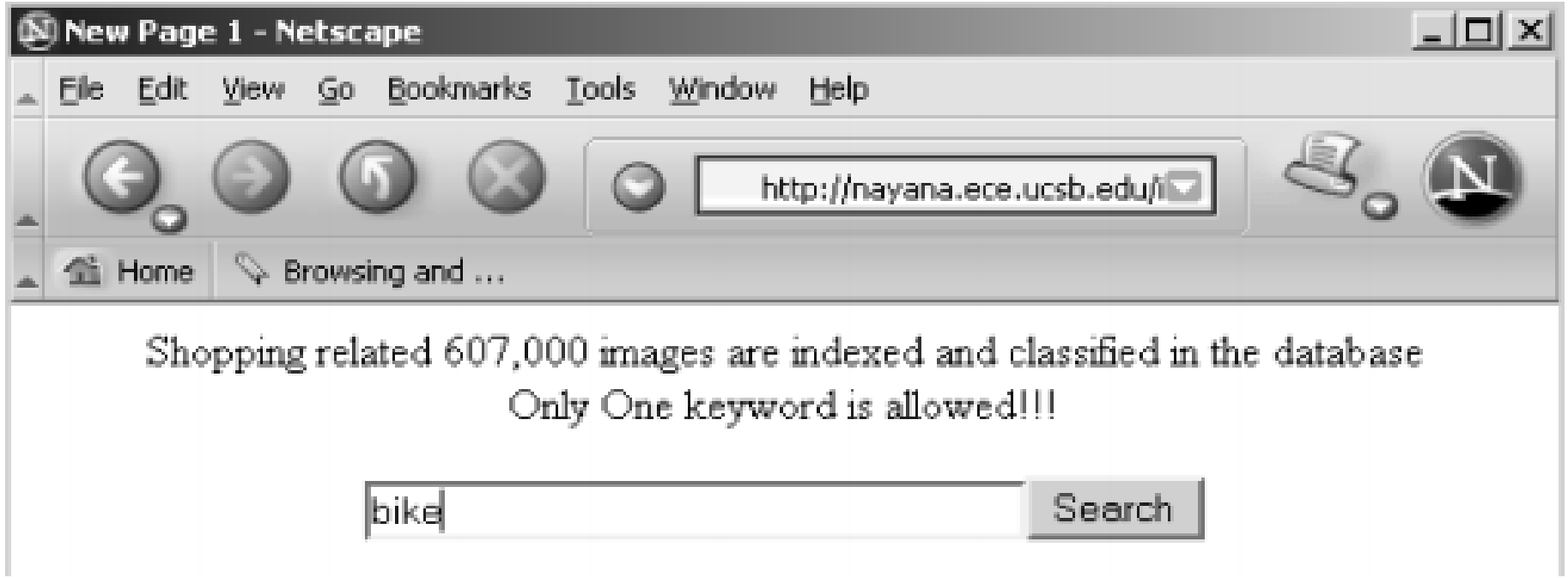
- Pengguna pada system implicit tidak menilai relevansi untuk kepentingan system IR, tetapi **hanya memuaskan kebutuhan mereka sendiri**.
- Pengguna pada system implicit tidak perlu diberitahu bahwa perilaku mereka (memilih dokumen) akan digunakan sebagai relevance feedback.

TOP K RETRIEVED (BLIND/PSEUDO RELEVANCE FEEDBACK)

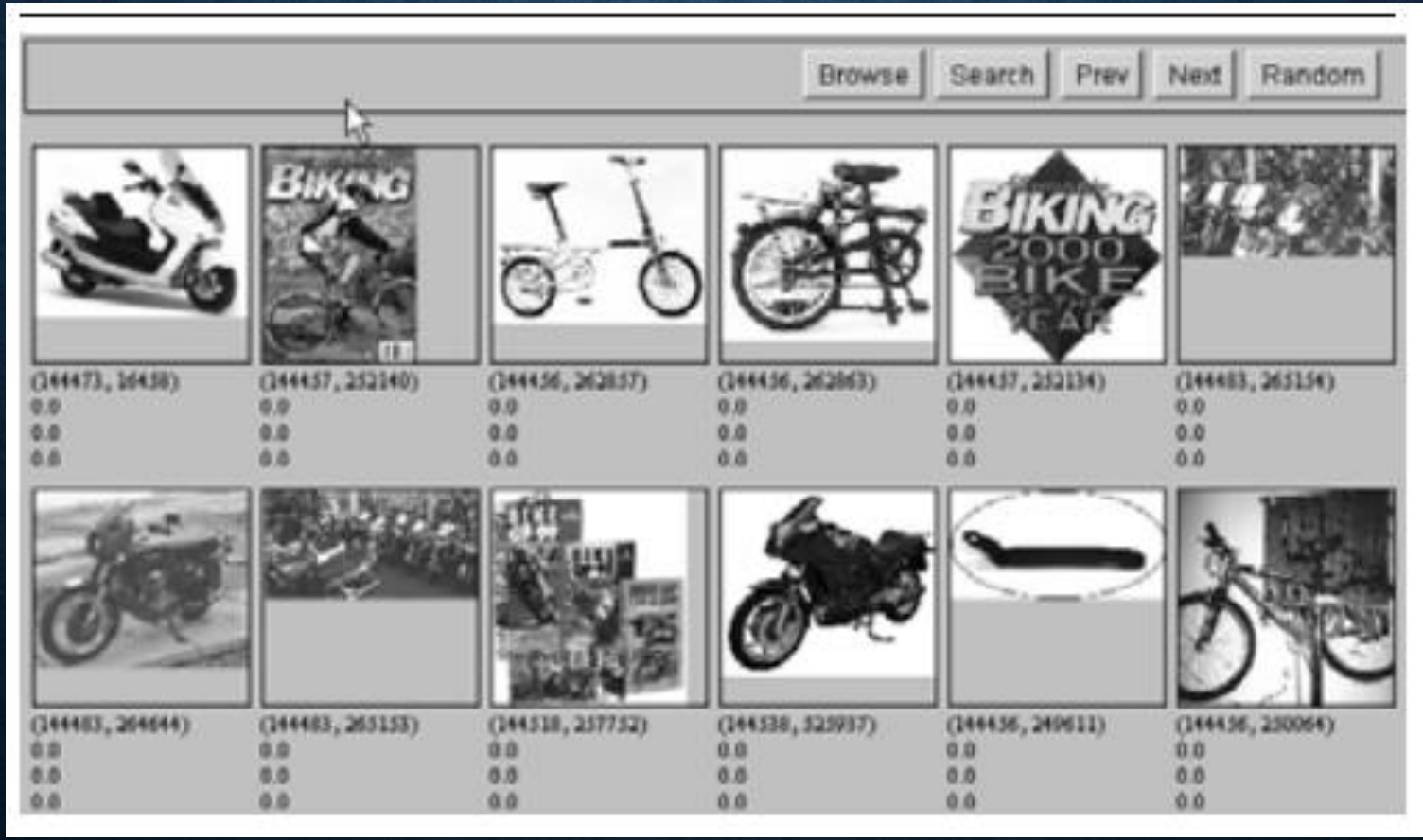
- Blind/pseudo relevance feedback tidak melibatkan user dalam perolehan keputusannya.
- Awal ide blind/pseudo relevance feedback adalah mengambil n dokumen teratas dan secara sederhana mengasumsikan dokumen tersebut relevan.
- Jika pengambilan n dokumen teratas dengan alasan yang cukup maka system akan mengambil query terms yang baik pula.
- Namun bila inisialisasi pengambil n dokumen tidak bagus akan terjadi pengambilan query atau terms yang buruk.

CONTOH

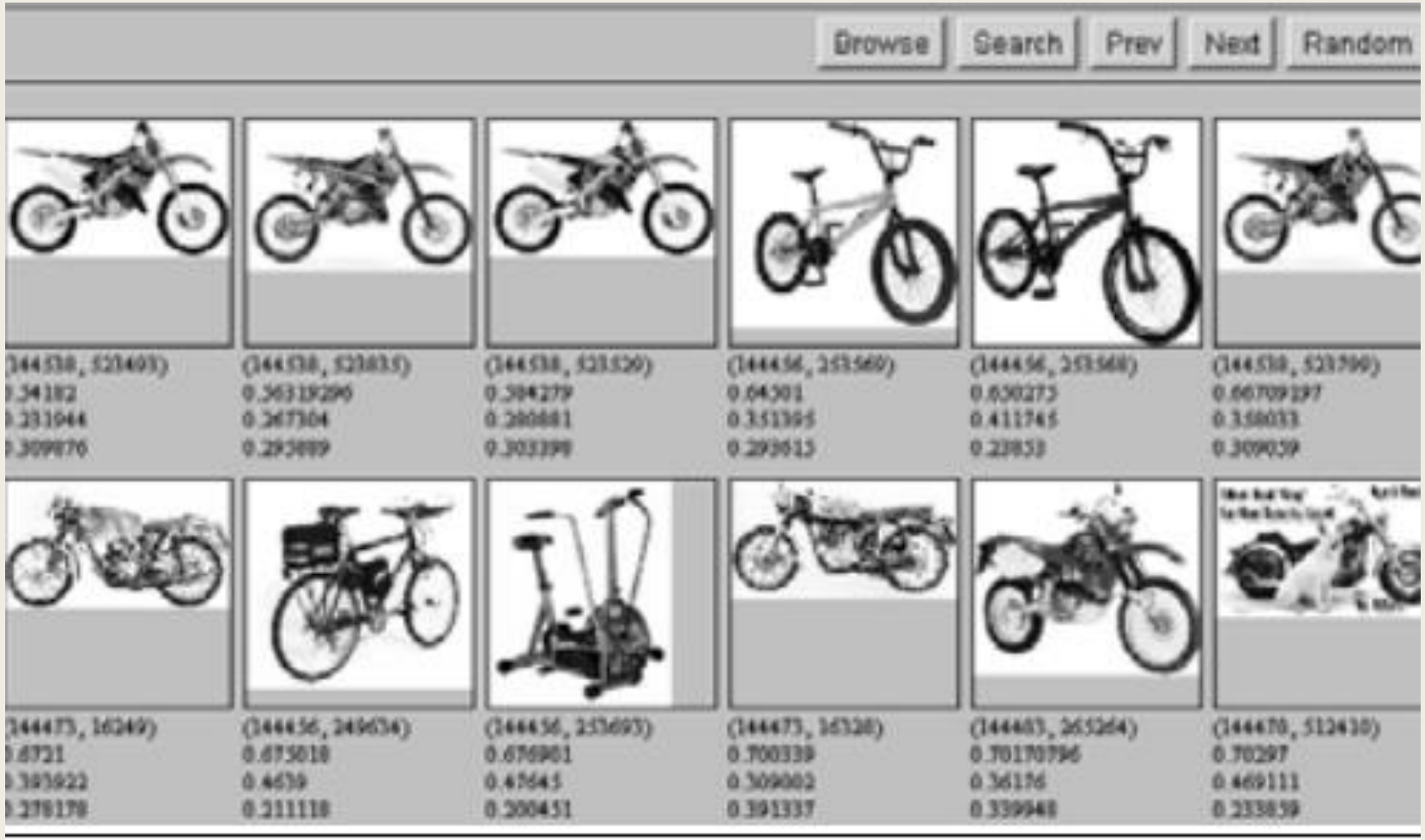
Image search engine
http://nayana.ece.ucsb.edu/imsearch/imsearch.html



HASIL QUERY AWAL



Hasil Setelah RF



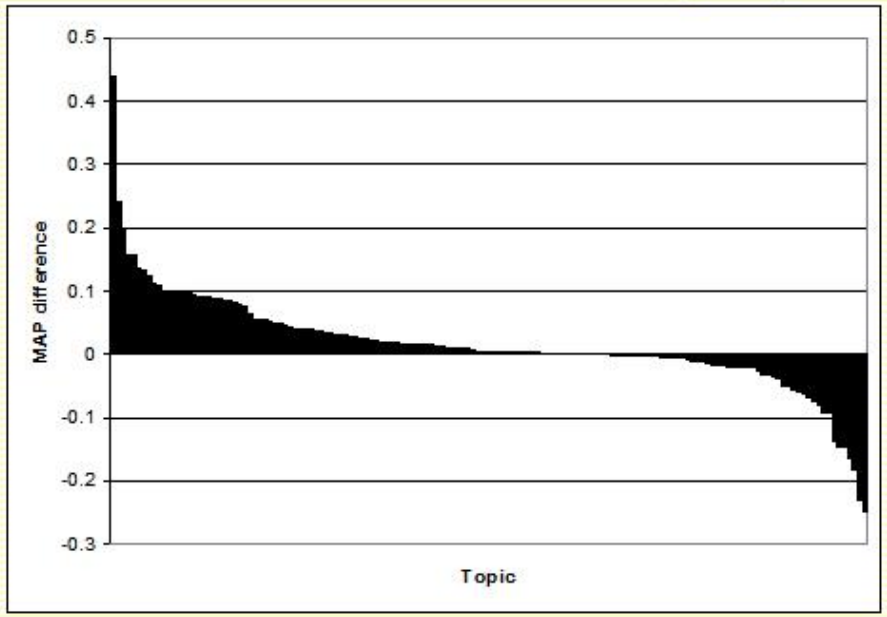
Contoh Lain

(Sumber gambar: Relevance Feedback
Jimmy Lin College of Information Studies
University of Maryland)

Results

	MAP	R-Precision
No feedback	0.1591	0.2022
With feedback	0.1806 (+13.5%)	0.2222 (+9.9%)

Blind relevance feedback doesn't always help!



Relevance Feedback

- Kita dapat mengubah query berdasarkan pada relevance feedback dan menerapkan vector space model.
- Gunakan hanya dokumen yang ditandai.
- Relevance feedback dapat meningkatkan recall dan precision.

Query expansion (QE)

- Query Expansion (QE) merupakan suatu teknik kembali informasi untuk memperbaiki query sehingga dapat memperoleh hasil yang lebih baik.
- Contoh: pada search engine Google kita mengetikan: tki temu kembali infromasi. Maka search engine melakukan query expansion (QE) dengan menampilkan saran: Did you mean: tki temu kembali informasi.

REFORMULASI QUERY

- Berdasarkan feedback dari user.
- Berdasarkan informasi yang diperoleh dari sekumpulan dokumen awal yang diperoleh.
- Berdasarkan pada informasi global dari koleksi dokumen.

Contoh QE

Google

tki temu kembali infromasi | Search

Advanced search

Everything

More

Show search tools

Did you mean: tki temu kembali *informasi*

[PDF] Materi UMUM kuliah **Temu Kembali Informasi (TKI)** - [Translate this page]
File Format: PDF/Adobe Acrobat - Quick View
Materi kuliah **Temu Kembali Informasi** (**TKI**) secara umum yaitu: 1. Materi UTS (Ujian Tengah Semester): -Pendahuluan (Introduction and Overview about ...
temukembaliinformasi.files.wordpress.com/.../materi-umum-kuliah-**temu-kembali-informasi-tki**1.pdf

Belajar **TKI (Temu Kembali Informasi)** - [Translate this page]
Bahan bacaan untuk belajar tentang Information Retrieval (IR) atau **Temu Kembali Informasi** (**TKI**), yaitu: Webpage: –http://www.cs.unt.edu/~rada/CSCE5200 ...
temukembaliinformasi.wordpress.com/ - Cached

Perbedaan & Persamaan Relevance Feedback (RF) dan Query Expansion (QE)

- Pada dasarnya kedua teknik ini (RF dan QE) tujuannya adalah untuk memperbaiki hasil temu kembali agar hasil dokumen/pencarian yang dilakukan menjadi lebih baik. Namun diantara keduanya memiliki perbedaan.
- Perbedaannya yaitu : Pada RF ada tanda/input/klik dari user untuk memperbaiki query sedangkan pada QE tidak ada (sistem secara automatis memperbaiki query).

Metode Reformulasi Query

1. **Global methods**

- QE menggunakan thesaurus atau WordNet
- QE melalui thesaurus otomatis
- Teknik mirip koreksi ejaan

2. **Local/basic methods**

- Relevance feedback
- Pseudo relevance feedback
- Indirect relevance feedback

Thesaurus

- Suatu thesaurus memberikan informasi tentang synonym dan kata-kata serta frase yang secara semantik berkaitan.

Misal (http://thesaurus.reference.com):

- **"market"**
- Part of Speech: verb
- Definition: package and sell goods
- Synonyms: advertise, barter, display, exchange, merchandise, offer for sale, retail, vend, wholesale.
- Antonyms: buy

Ekspansi Query Dengan Thesaurus

- Tidak memerlukan input dari user.
- Untuk setiap kata t pada suatu query, ekspansi query dengan sinonim dan kata lain t dari thesaurus.
- Bobot kata-kata tambahan dapat lebih kecil daripada kata-kata pada query awal.
- Biasanya meningkatkan recall.
- Banyak digunakan pada bidang ilmu pengetahuan / teknik.

WORDNET

- http://www.cogsci.princeton.edu/~wn/
- Suatu database yang detil berisi hubungan semantik antara kata- kata dalam bahasa Inggris.
- Kira- kira berisi 144,000 kata dalam bahasa Inggris.
- Kata benda, sifat, kerja, dan keterangan dikelompokkan menjadi 109,000 set sinonim yang disebut synsets.

Hubungan Pada WordNet Synset

- Antonym : front → back
- Attribute : benevolence → good (noun to adjective)
- Pertainym : alphabetical → alphabet (adjective to noun)
- Similar : unquestioning → absolute
- Cause : kill → die
- Holonym : chapter → text (part-of)
- Meronym : computer → cpu (whole-of)
- Hyponym : tree → plant (specialization)
- Hypernym : fruit → apple (generalization)

QE menggunakan WordNet

- Tambahkan sinonim pada synset yang sama.
- Tambahkan hiponim untuk memasukkan katakata khusus.
- Tambahkan hipernim untuk membuat query lebih umum.
- Tambahkan kata-kata lain yang berkaitan untuk memperluas query.

QE menggunakan WordNet

- Contoh query awal :

  "information system"

- WordNet (synonym):

1. information : message, content, subject matter, substance
2. system : group, grouping

- Query expansion:

  information message system group

Tipe Ekspansi Query

Global Analysis: (statis; dari semua dokumen dalam koleksi)

- Controlled vocabulary
- Manual thesaurus
- Automatically derived thesaurus (kemunculan secara statistik)
- Based on query log mining (umum di web)

Local Analysis: (dynamic)

- Analisis dokumen yang terambil

PSEUDO RELEVANCE FEEDBACK

Metode untuk analisis lokal secara otomatis :

- Menggunakan metode relevance feedback tanpa input eksplisit dari user.
- Hanya asumsikan dokumen yang diperoleh pada top n adalah relevan, dan gunakan untuk membentuk query yang baru.
- Query expansion diperbolehkan berisi kata-kata yang berkaitan dengan kata-kata pada query.

Pseudo Relevance Feedback

- Ambil top n dokumen
- Dari semua kata-kata pada dokumen tsb.,

  ambil top t kata
- Urutan kata-kata menunjukkan cara kata-kata tersebut diurutkan :

  n (banyaknya dokumen yang berisi kata t)

  f (jumlah kemunculan kata t)

  n * idf

  f * idf

Pseudo Relevance Feedback

- Contoh: Top 3 dokumen:
  - D1 : A, B, B, C, D
  - D2 : C, D, E, E, A, A
  - D3 : A, A, A
  - Asumsikan idf dari A=1, B=1, C = 1, D=2, E = 2
- Rank:

kata	n	f	n * idf	f * idf
A	3	6	3	6
B	1	2	1	2
C	2	2	2	2
D	2	2	4	4
E	1	2	4	8

The Rocchio algorithm for relevance feedback

- Rocchio classifiers merupakan salah satu metode pembelajaran supervised document classification.
- Metode klsifikasi rocchio membandingkan kesamaan isi antara data training dan data test dengan merepresentsaikan semua data ke dalam sebuah vector.
- Kedekatan kesamaan isi dihitung dari kedekatan sudut yang terbentuk antara bobot data training dan bobot data test menggunakan aturan sodinus.
- Untuk menghitung bobot setiap kata dalam dokumen digunakan skema pembobotan tfidf (Term Frequency / Invers Document Frequency) karena komponen heuristic utama adalah klasifikasi rocchio yaitu skema pembobtan tfidf, metode pembelajaran rocchio disebut juga dengan tfidf Classifiers (Joachihms, 1997)

Langkah-langkah pengolahan query

- **Text Mining dan Klasifikasi Teks** : mencari dan mengelompokkan dokumen ke dalam kategori tertentu
- **Parsering** : memilah isi dokumen menjadi unit-unit kecil (token), yang berupa kata, frase, atau kalimat.
- **Stemming** : proses penghilangan prefiks (awalan) dan sufiks (akhiran)
- **Inverted Index** : struktur yang dioptimasi untuk menemukan kembali dokumen

Keterbatasan Teknik Rocchio

- Teknik Rocchio ini memiliki keterbatasan dimana sering gagal untuk mengklasifikasikan kelas multimodal dan hubungan.
- Misalnya, negara Burma berganti nama menjadi Myanmar pada tahun 1989. Oleh karena itu dua pertanyaan dari "Burma" dan "Myanmar" akan muncul lebih jauh terpisah dalam vector space model, meskipun mereka sama-sama mengandung asal yang sama.

CONTOH PENGGUNAAN ROCCHIO ALGORITHM

- **Query = “Sistem”**
- Dokumen 1 : “Sistem Adalah Kumpulan Elemen”
- Dokumen 2 : “Adalah Kumpulan Elemen Yang Saling Berinteraksi”
- Dokumen 3 : “Sistem Berinteraksi Untuk Mencapai Tujuan”

Contoh Penggunaan Rocchio Algorithm

Proses parsering

Doc 1 : sistem, adalah, kumpulan, elemen

Doc 2 : adalah, kumpulan, elemen, yang, saling, berinteraksi

Doc 3 : sistem, berinteraksi, untuk, mencapai, tujuan

Proses filtering

Doc 1 : sistem, kumpulan, elemen

Doc 2 : kumpulan, elemen, saling, berinteraksi

Doc 3 : sistem, berinteraksi, mencapai, tujuan

Proses Stemming

Doc 1 : sistem, kumpul, elemen

Doc 2 : kumpul, elemen, saling, interaksi

Doc 3 : sistem, interaksi, capai, tujuan

Menghitung tf-idf

term	tf				df	D/df	Log (D/df)	W			
	Q	D1	D2	D3				Q	D1	D2	D3
Capai	0	0	0	1	1	3	0,477	0	0	0	0,477
Elemen	0	1	1	0	2	1,5	0,176	0	0,176	0,176	0
Interaksi	0	0	1	1	2	1,5	0,176	0	0	0,176	0,176
Kumpul	0	1	1	0	2	1,5	0,176	0	0,176	0,176	0
Saling	0	0	1	0	1	3	0,477	0	0	0,477	0
Sistem	1	1	0	1	2	1,5	0,176	0,176	0,176	0	0,176
tujuan	0	0	0	1	1	3	0,477	0	0	0	0,477

MENGHITUNG JARAK DOKUMEN DAN QUERY

Term	W(Q)*W(Di)			Panjang Vektor				
	D1	D2	D3	Q^2	D1^2	D2^2	D3^2	
Capai	0	0	0	0	0	0	0,227	
Elemen	0	0	0	0	0,031	0,031	0	
Interaksi	0	0	0	0	0	0,031	0,031	
Kumpul	0	0	0	0	0,031	0,031	0	
Saling	0	0	0	0	0	0,227	0	
Sistem	0,031	0	0,031	0,031	0,031	0	0,031	
Tujuan	0	0	0	0	0	0	0,227	
	0,031	0	0,031	0,031	0,093	0,321	0,517	SUM
				0,176	0,305	0,566	0,719	SQRT

COSINE SIMILARITY

Hitung kemiripan Q dengan D1, D2 dan D3

D1	D2	D3
0,57	0	0,25

Urutkan hasil perhitungan kemiripan

D1	D3	D2
0,57	0,25	0