Data Mining

Pertemuan 10: Web Mining

Outline Materi

- Web Content Mining
- Web Usage Mining
- Web Structure Mining

Web Mining

- Web Data Mining:
  - Teknik untuk menemukan dan mengekstraksi informasi dari dokumen atau layanan yang ada di Web secara otomatis
- Riset Web Mining meliputi:
  - Database
  - Information Retrieval
  - Machine Learning
  - Natural Language Processing

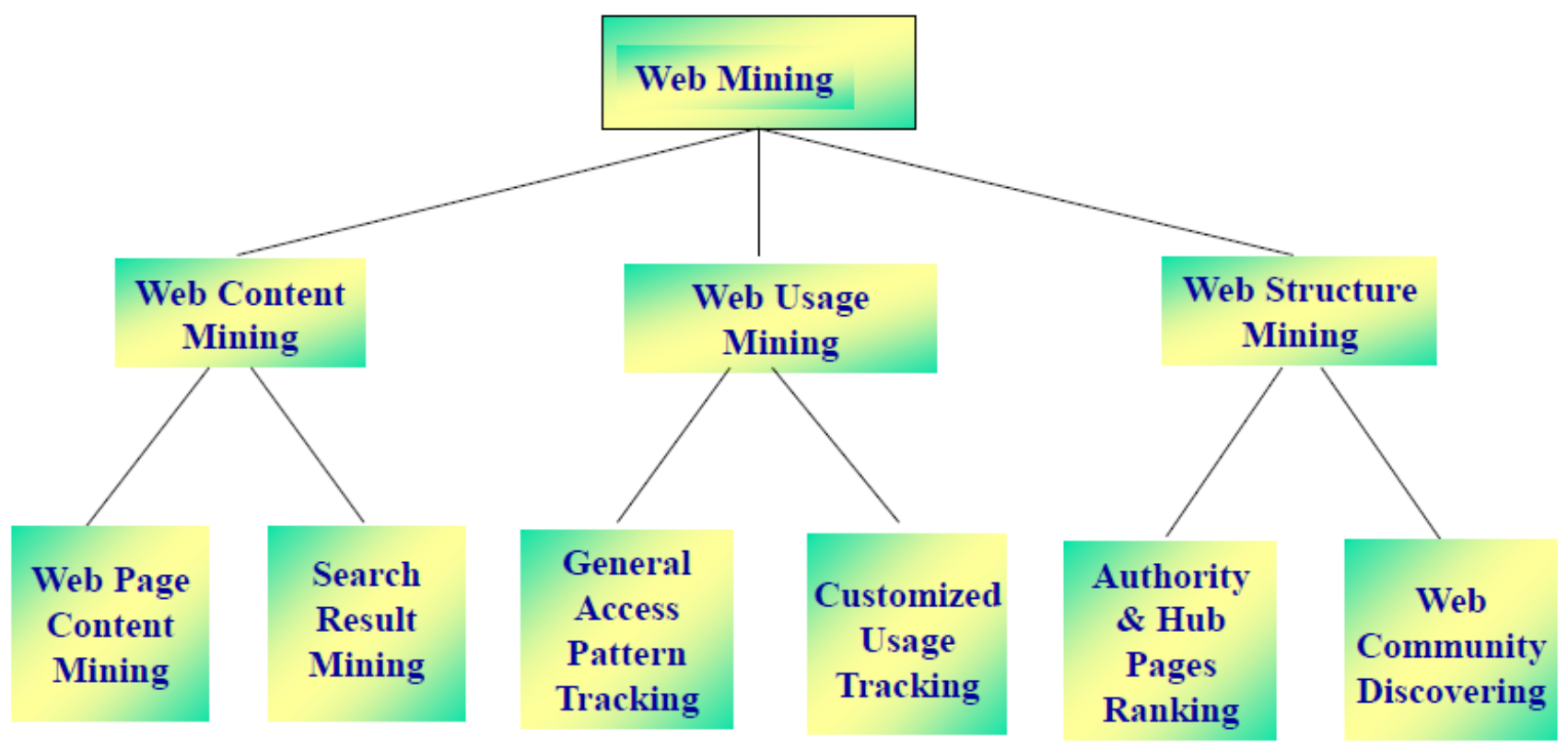
Mining pada Web

- Web adalah pusat informasi untuk:
  - Layanan informasi, seperti:
    - Berita, iklan, informasi konsumen, manajemen keuangan, pendidikan, pemerintahan, e-commerce, dll
  - Informasi hyperlink
  - Perilaku pengguna (informasi akses dan penggunaan)
  - Isi dan pengorganisasian website
  - Media sosial

Tantangan dalam Web Mining

- Mencari:
  - Keteraturan (*regularity*) dan dinamika dari isi Web
  - Pola akses pengguna Web
  - Struktur Web
- Permasalahan:
  - Jumlah permasalahan yang sangat banyak (data banyak namun miskin informasi)
  - Keterbatasan cakupan Web: halaman Web tersembunyi, sebagian besar data ada di DBMS
  - Dinamik dan semi-terstruktur
  - Keterbatasan pada pencarian yang berorientasi pada kata kunci
  - Keterbatasan pada kustomisasi masing-masing individu pengguna

Taksonomi Web Mining



1. Web Content Mining

- Penemuan informasi bermanfaat dari isi/data/dokumen dari Web, mencakup data teks, gambar, video, audio, metadata, dan hyperlink
- Cara pandang *information retrieval* (terstruktur dan semi-terstruktur):
  - Membantu/meningkatkan pencarian informasi
  - Penyaringan informasi berdasarkan profil pengguna
  - Ekstraksi informasi

Permasalahan terkait Web Content Mining

- Pembuatan *tool* cerdas untuk Information Retrieval:
  - Pencarian kata kunci dan frasa kunci
  - Penemuan aturan tata bahasa dan kolokasi
  - Klasifikasi hypertext
  - Ekstraksi frasa kunci dari dokumen teks
  - Pembuatan model/aturan pembelajaran untuk ekstraksi
  - Hierarchical Clustering
  - Prediksi keterkaitan antar-kata

Implementasi

- Penyaringan Informasi/Pengkategorian
  - Collaborative Filtering
- *Personalized Web Agents*
  - Web Wrapper

Penyaringan Informasi/Pengkategorian

- Menggunakan berbagai macam teknik *information retrieval* dan karakteristik dokumen hypertext pada Web untuk mengambil, menyaring, dan mengkategorikan dokumen secara otomatis
  - **HyPursuit**: menggunakan informasi semantik yang ada pada struktur link dan isi dokumen untuk membuat hirarki cluster dari dokumen hypertext dan membuat struktur dari domain informasi
  - **BO (Bookmark Organizer)**: Mengkombinasikan teknik *hierarchical clustering* dan interaksi pengguna untuk mengatur sekumpulan dokumen Web berdasarkan informasi konseptual

HyPursuit: fungsi similaritas dari halaman web dengan menggunakan hyperlink

- Kesamaan hyperlink dari 2 dokumen Web, *i* dan *j*, dapat dituliskan dalam persamaan:

$$S_{ij} = W_d \bullet S_{ij}^{dec} + W_a \bullet S_{ij}^{anc} + W_s \bullet S_{ij}^{spl}$$

Where Common Descendants: $S_{ij}^{dec} = \sum_{x \in_{decendants}^{common}} \frac{1}{2^{(spl_{ix}^{\vec{ij}} + spl_{jx}^{\vec{ii}})}}$

Common Ancestors: $S_{ij}^{anc} = \sum_{x \in_{ancestors}^{common}} \frac{1}{2^{(spl_{xi}^{\vec{ij}} + spl_{xj}^{\vec{ii}})}}$

Shortest path length between documents: $S_{ij}^{spl} = \frac{1}{2^{(spl_{ij})}} + \frac{1}{2^{(spl_{ji})}}$

W_d, W_a, and W_s are damping factors for normalization.

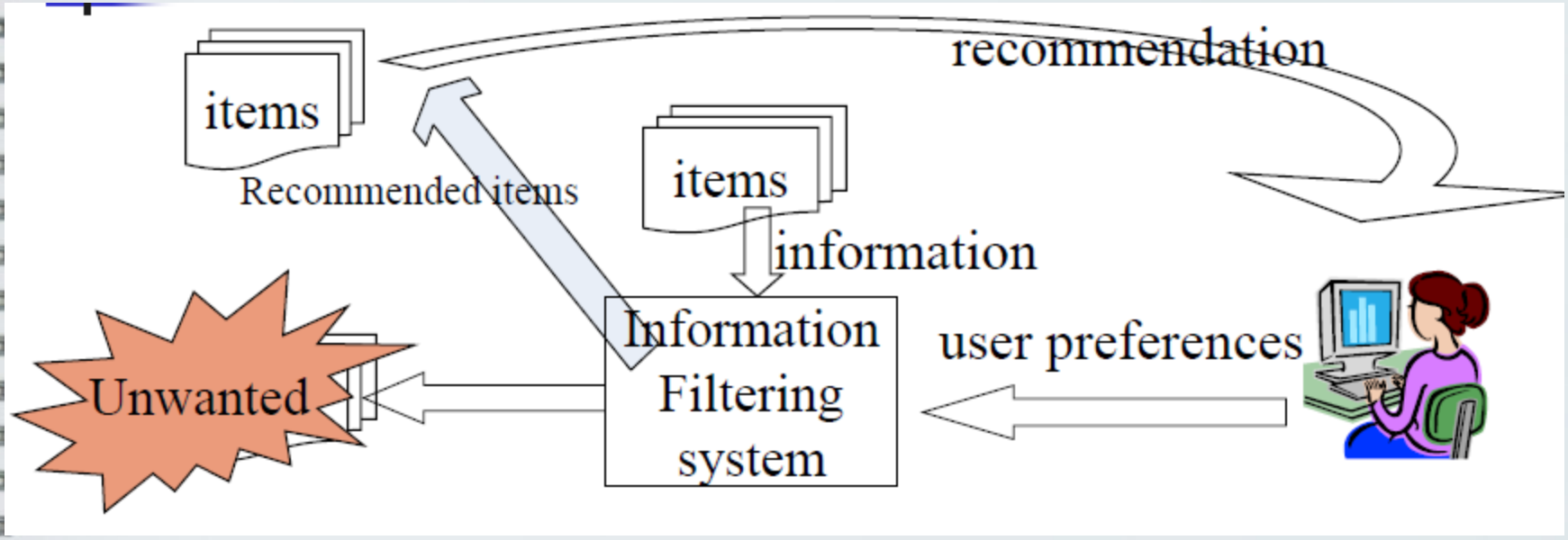
spl_{xy} ≡ length of a shortest path between d_x and d_y.

spl_{xy}^{z} ≡ length of a shortest path between d_x and d_y not travelling d_z

Bagaimana cara mencari laman Web yang mirip?

- Pendekatan berdasarkan isi
- Pendekatan berdasarkan struktur
- Gabungan antara pendekatan isi dan struktur

Penyaringan Informasi

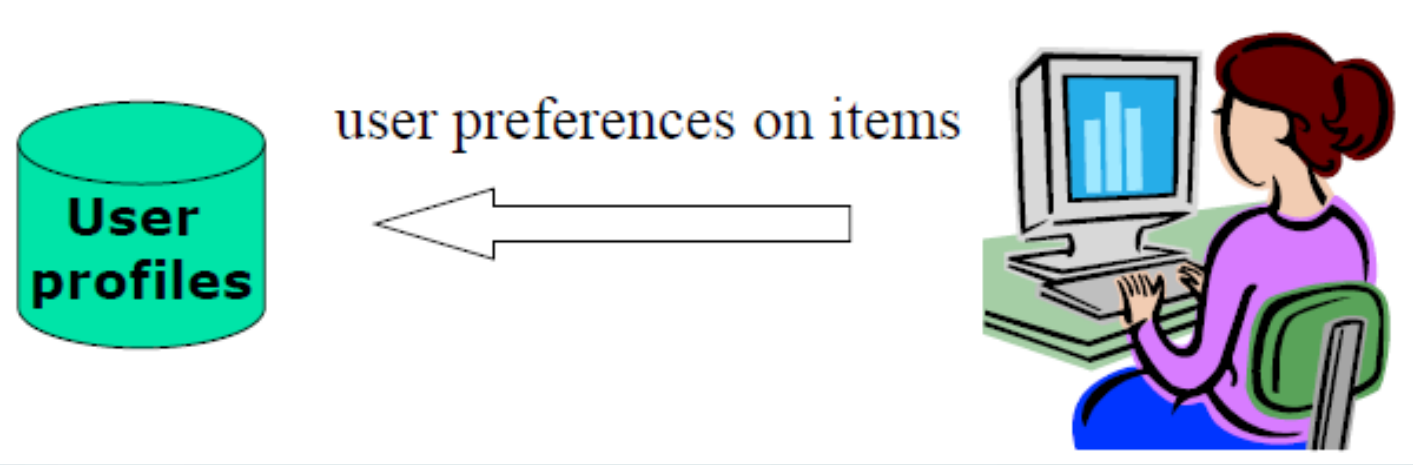


Mengapa perlu ada penyaringan informasi?

- Menghemat waktu
- Mencari item yang dianggap “menarik”

Asumsi

- Data preferensi pengguna dapat disimpan

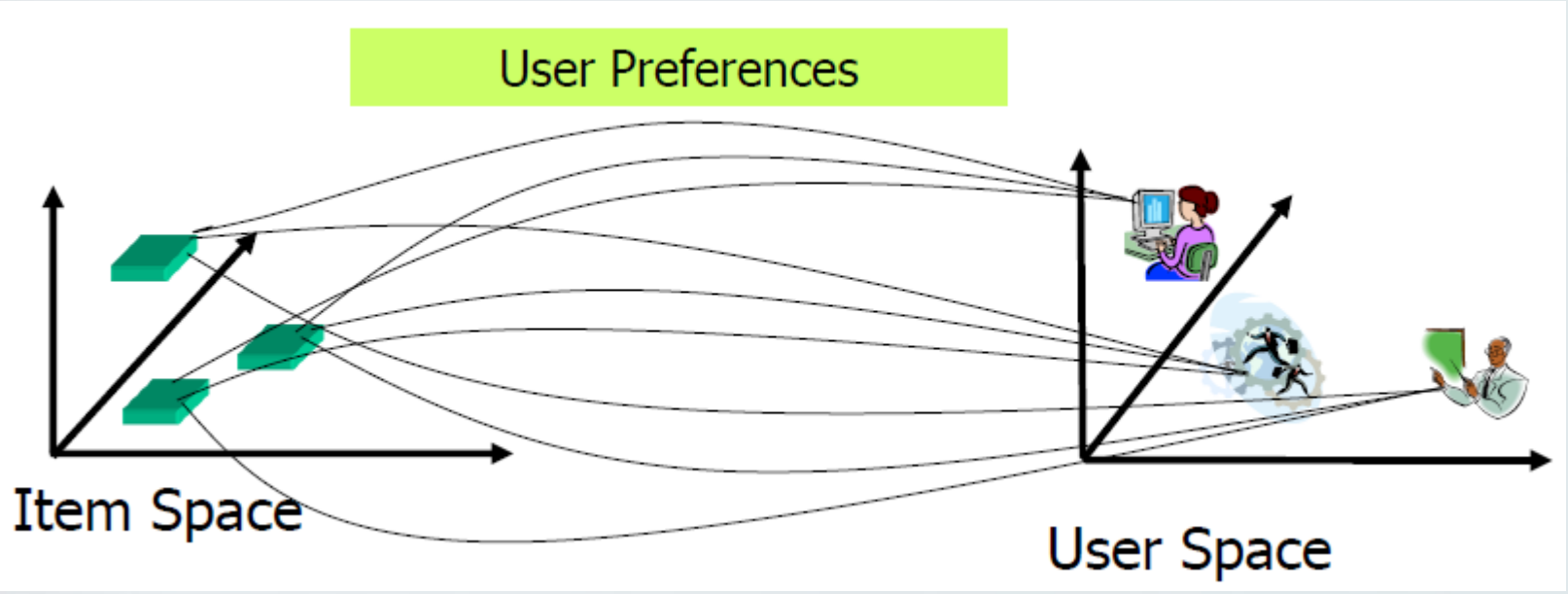


Definisi dari Permasalahan Penyaringan Informasi

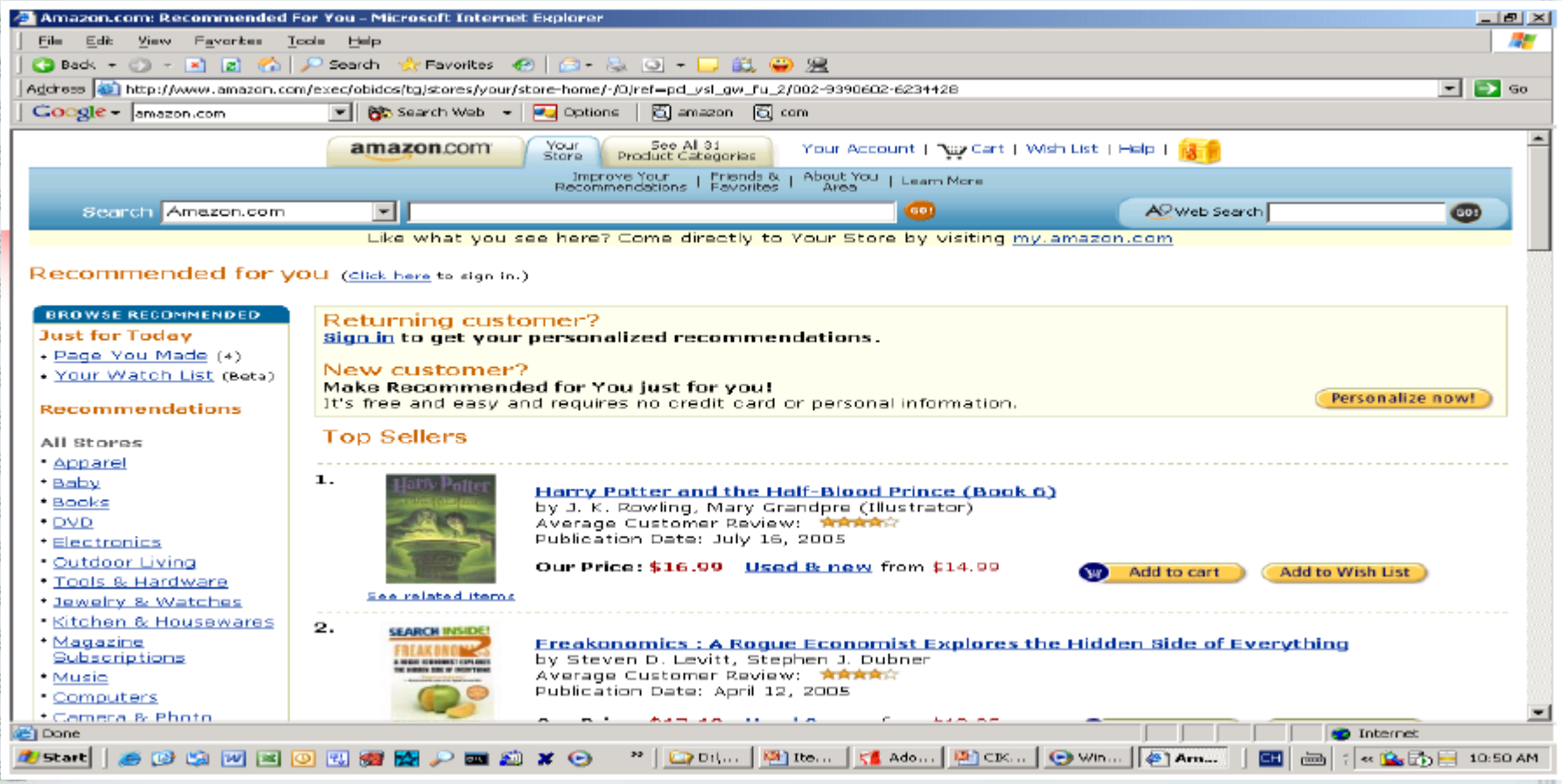
- Diberikan sebuah dataset D sebagai himpunan $<U_i, I_j, O_{ij}>$ dimana:
  - U_i: user ke-i dalam sistem
  - I_j: Item ke-j dalam sistem
  - O_{ij}: opini user ke-i terhadap item j
- Carilah sejumlah *k* item rekomendasi kepada user

Mapping dari dua ruang dimensi

- Q1: Untuk sebuah item, tipe pelanggan seperti apa yang akan menyukainya?
- Q2: Untuk seorang pelanggan, item seperti apa yang akan disukainya?



Aplikasi dari Penyaringan Informasi: e-Commerce



Tantangan Penyaringan Informasi

- Ketepatan prediksi
- **Skalabilitas**: jika jumlah pengguna dan item meningkat secara signifikan, bagaimana kinerja algoritma?
- **Kehandalan**: jika terdapat *noise* pada data, bagaimana algoritma dapat memberikan prediksi yang tepat?
- Renggangnya ruang data: matriks pemetaan pengguna-item sangat renggang (banyak mengandung nilai 0)
- ***Cold Start**:* Bagaimana membuat rekomendasi bagi pengguna baru atau item baru?

2. Web Usage Mining

- Web Log Mining
  - Pre-processing
  - Pattern mining
  - Pattern analysis

Sumber: https://d279iyy6fmg6l4.cloudfront.net/blog/log-mining.png

Aplikasi

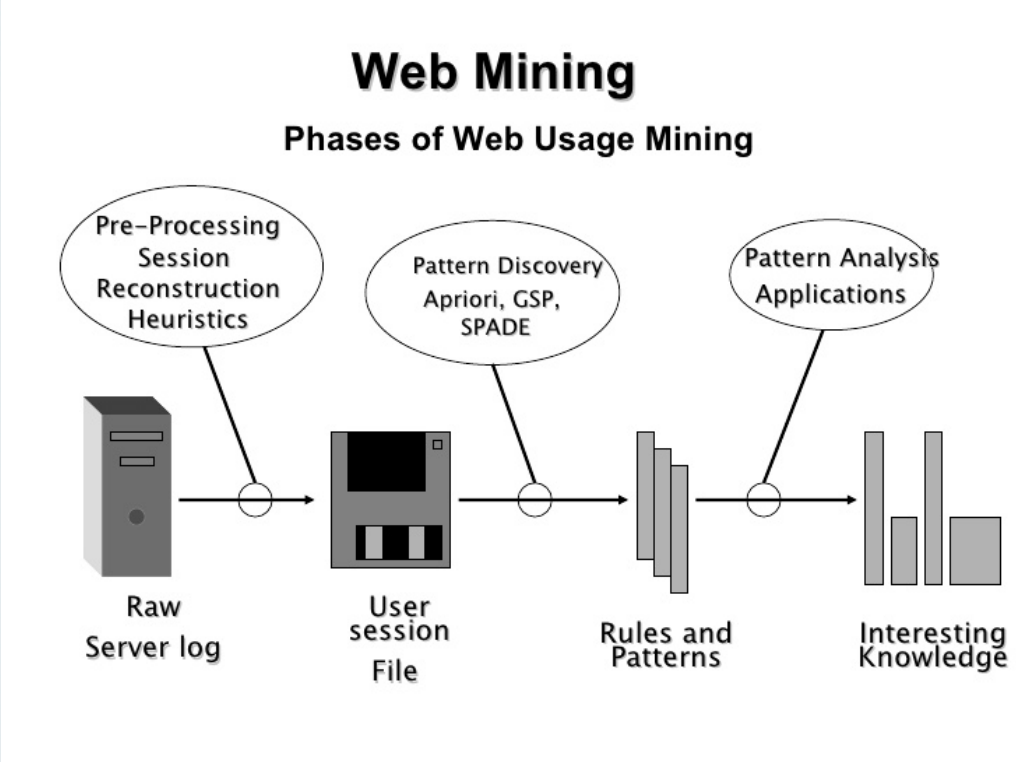
- Menyasar pelanggan yang berpotensi dalam e-commerce
- Meningkatan kualitas dan kuantitas layanan informasi dari internet
- Meningkatkan kinerja web server (*Load Balancing*)
- Mengidentifikasi lokasi yang tepat untuk memasang iklan
- Meningkatkan desain sebuah situs web
- *Fraud/intrusion detection*
- Memperkirakan aksi pengguna (yang memungkinan adanya *pre-fetching*)

Outcome

- Aturan asosiasi
  - Mencari halaman Web yang sering dilihat bersamaan
- Clustering
  - Mengelompokkan pengguna berdasarkan pola *browsing*
  - Mengelompokkan halaman Web berdasarkan isinya
- Klasifikasi
  - Merelasikan atribut pengguna terhadap pola

Fase

- Pre-processing
- Penemuan pola
- Analisis pola



Sumber: http://pubs.sciepub.com/ajss/3/2/3/figure/5

Fase 1: Pre-processing

- Konversi data mentah ke abstraksi data yang diperlukan untuk penerapan algoritma data mining:
  - Memetakan data log ke tabel relasional sebelum menerapkan metode data mining yang sudah diadaptasi
  - Menggunakan data log secara langsung dengan menerapkan sejumlah teknik pre-processing

Data mentah – log Web

- **Click stream:** serangkaian permintaan untuk membuka halaman
- **User session:** Serangkaian klik user terhadap satu atau beberapa Web server
- **Server session**: sekumpulan klik user terhadap sebuah Web Server selama user session tertentu
- **Episode:** Sub-himpunan dari klik user yang saling berkaitan yang muncul dalam sebuah user session tertentu

Fase 2: Penemuan Pola

- Penemuan pola menggunakan sejumlah teknik seperti analisis statistik, aturan asosiasi, clustering, klasifikasi, *sequential pattern*, *dependency modeling*

Fase 3: Analisis Pola

- Sebuah proses untuk mendapatkan pengetahuan tentang bagaimana para pengunjung menggunakan Website agar developer dapat:
  - Mencegah disorientasi dan membantu desainer web untuk meletakkan informasi/fungsi penting tepat dimana pengunjung akan melihat dan bagaimana cara penggunaannya
  - Membangun sebuah Web server yang adaptif

3. Web Structure Mining

- Tujuannya adalah untuk **menemukan struktur link** dari hyperlink pada level inter-dokumen untuk menghasilkan ringkasan terstruktur tentang situs dan laman Web:
  - Arah 1: Berdasarkan hyperlink, **mengkategorikan laman Web** dan menghasilkan informasi
  - Arah 2: Menemukan **struktur dari sebuah dokumen Web**
  - Arah 3: Menemukan inti dari hirarki jaringan antar-link pada **laman Web dari sebuah domain tertentu**

Aplikasi

- Pengkategorian/perankingan laman Web
- Penemuan komunitas-komunitas baru
- Penemuan skema pada lingkungan semi-terstruktur

Metode Populer

- HITS (distilasi topik)
- PageRank (Perankingan laman Web yang digunakan oleh Google)
- Algoritma oleh komunitas cyber

HITS:
Hyperlink Induced Topic Search

- Memandang Web sebagai sebuah graph berarah
- Asumsi: jika dokumen A memiliki hyperlink ke dokumen B, maka penulis dokumen A berpendapat bahwa dokumen B memiliki informasi yang berguna
- Berhubungan dengan pengidentifikasian laman Web mana yang memiliki otoritas paling tinggi dalam lingkup topik yang luas
- Merupakan komputasi berbasis link semata, tanpa melihat keterkaitan konteks

HITS: Hub dan Otoritas

- Menentukan 2 nilai untuk sebuah laman web:
  - **Hub:** Nilai dari semua linknya terhadap laman lain
  - **Otoritas:** Nilai dari isi laman tersebut
- **Mutual Reinforcing Relationship**
  - Nilai otoritas dihitung sebagai jumlah dari nilai hub yang menuju laman tersebut
  - Nilai hub adalah jumlah nilai otoritas yang ditunjuk oleh laman tersebut
  - Otoritas yang baik adalah laman ditunjuk oleh banyak hub yang baik,
  - Sedangkan hub yang baik adalah laman yang menunjuk banyak otoritas yang baik

HITS: Algoritma

- Pengambilan sampel, yang terdiri atas beberapa ribu laman Web dari sebuah hasil query
- Penentuan komponen bobot propagasi, yang mengestimasi hub dan otoritas secara iteratif:
  - Update otoritas: update skor Otoritas tiap node
  - Update hub: update skor Hub tiap node
- Hasilnya, laman yang memiliki bobot tertinggi akan dikembalikan sebagai laman hub dan otoritas dari topik hasil query

Keterbatasan HITS

- Terbatas hanya pada cakupan topik yang sempit
  - Jumlah laman yang otoritatif kurang
  - Penambahan sedikit edge dapat berpotensi mengubah skor secara besar-besaran
- *Topic drifting*, muncul ketika laman hub mendiskusikan banyak topik

PageRank

- Diperkenalkan oleh Brin dan Page (1998)
  - Menambang struktur hyperlink Web untuk menghasilkan ranking kepentingan yang bersifat “global” dari semua laman Web
- Asumsi: Laman yang banyak di-link lebih “penting” dari yang sedikit di-link
- Sebuah laman memiliki ranking tinggi jika jumlah ranking dari *back-link*-nya juga tinggi
- Google menggunakan beberapa faktor untuk meranking hasil pencarian, antara lain: kedekatan lokasi, *anchor text*, PageRank, query, dll

PageRank: Ide Dasar

- Merupakan hasil voting dari semua laman Web tentang seberapa penting sebuah laman Web tertentu
- Hyperlink terhadap sebuah laman dihitung sebagai sebuah dukungan voting
- PageRank dari sebuah laman didefinisikan secara rekursif dan bergantung kepada jumlah dan metrik PageRank dari semua laman yang me-*link*-nya
- Laman yang di-*link* oleh banyak laman dengan PageRank yang tinggi juga akan mendapat ranking yang tinggi
- Jika tidak ada link ke laman web, maka tidak ada support/dukungan untuk laman tersebut

PageRank: Algoritma

- Menggunakan distribusi probabilitas untuk merepresentasikan *likelihood/*kecenderungan dari seseorang yang meng-klik link akan tiba pada sembarang laman
- Mengasumsikan bahwa distribusinya tersebar secara merata di antara seluruh koleksi dokumen
- Komputasi PageRank membutuhkan beberapa iterasi di dalam koleksi dokumen untuk menyesuaikan nilai PageRank

HITS vs PageRank

- Keduanya merupakan algoritma iteratif yang didasarkan pada link pada dokumen Web
- HITS dieksekusi saat query, sedangkan PageRank dijalankan saat *indexing*
- HITS jarang digunakan oleh mesin pencari
- HITS menghitung 2 skor untuk sebuah laman, hub dan otoritas, sedangkan PageRank menghasilkan 1 skor
- HITS diterapkan pada sebuah himpunan kecil dari dokumen yang relevan, sedangkan PageRank diterapkan pada semua laman Web

3. Komunitas Web

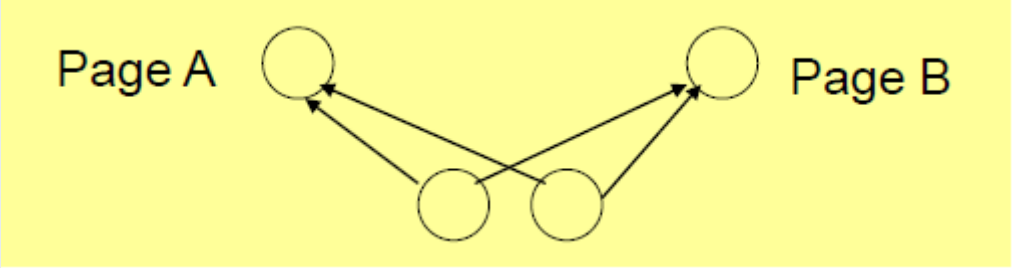
- Komunitas cyber adalah sekumpulan laman Web yang memiliki kesamaan ketertarikan *(interest)*
- Sifat utama:
  - Laman-laman pada komunitas yang sama akan mirip secara isi
  - Laman dalam satu komunitas akan berbeda dengan laman dalam komunitas lain
  - Mirip dengan cluster

Penemuan Komunitas

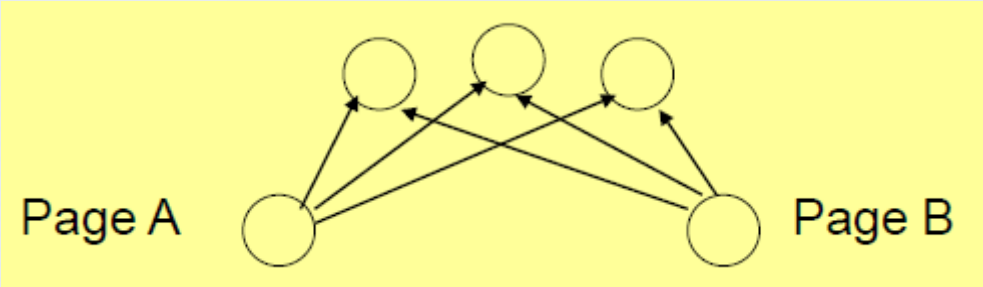
- Penemuan komunitas Web mirip dengan proses clustering
- Perlu didefinisikan similaritas antara dua laman Web

Similaritas Laman Web

- **Co-citation:** similaritas A dan B diukur dengan jumlah laman yang mensitasi A dan juga B



- **Bibliographic coupling:** similaritas A dan B diukur dengan jumlah laman yang disitasi oleh A dan juga B

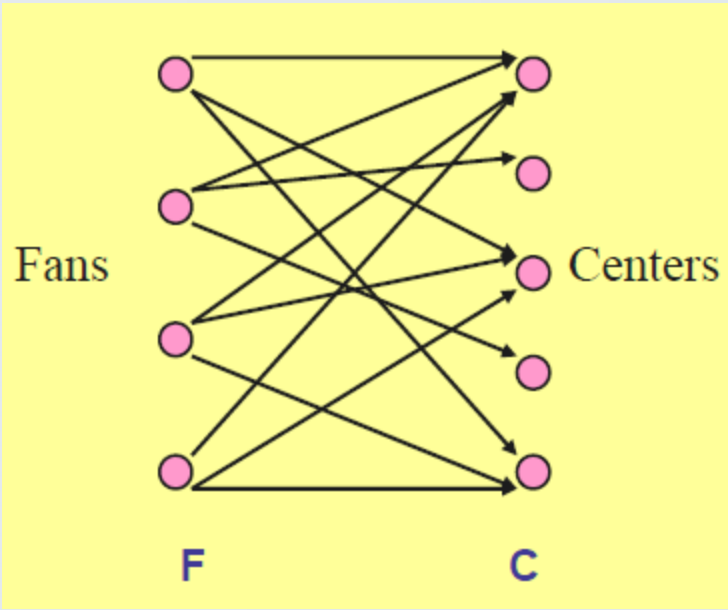


Algoritma CT

- Metode dari IBM Almaden Research Center, Clever search engine
- Metode ini disebut *Community Trawling (CT)*
- Diterapkan pada graph dari 200 juta laman Web dan bekerja dengan sangat baik

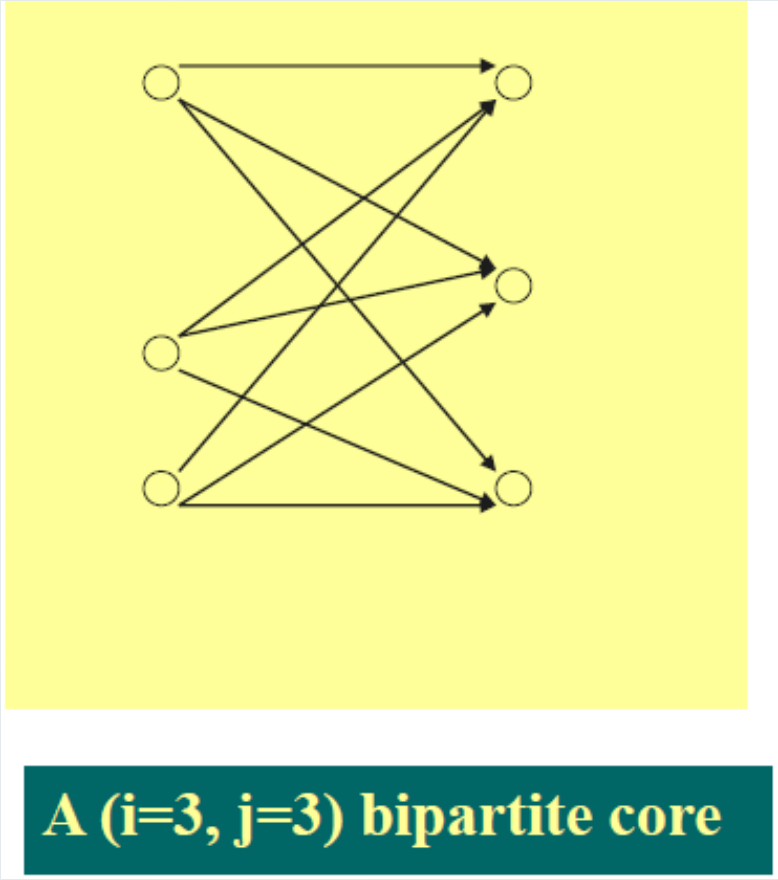
CT: Ide Dasar

- Definisi dari komunitas Web
  - Adanya sub-graph berarah yang bipartit dan padat
- Graph Bipartit: Node dapat dibagi menjadi 2 himpunan, F dan C
  - Setiap edge berasal dari F dan mengarah ke C
  - Graph yang padat jika terdapat banyak edge dari F ke C



Inti Bipartit

- Inti Bipartit:
  - Sebuah graph bipartit lengkap dengan setidaknya terdapat i node dari F dan j node dari C
  - i dan j adalah parameter yang dapat dikustomisasi
- Setiap komunitas memiliki inti bipartit dengan nilai i dan j tertentu



A (i=3, j=3) bipartite core

Algoritma CT

- Adanya sebuah inti bipartit adalah identitas dari sebuah komunitas
- Untuk mengekstraksi semua komunitas berarti mengidentifikasi semua inti bipartit pada web
- Penulis algoritma ini mendapatkan algoritma yang efisien untuk mengidentifikasi semua inti bipartit dengan pendekatan *iterative pruning*
  - Elimination-generation pruning

Kekurangan CT

- Graph bipartit tidak dapat mengakomodir semua jenis komunitas
- Kerapatan dari komunitas merupakan hal yang sangat sulit untuk disesuaikan

Ringkasan

- Web Mining
  - Content Mining
  - Usage Mining
  - Structure Mining