Text mining

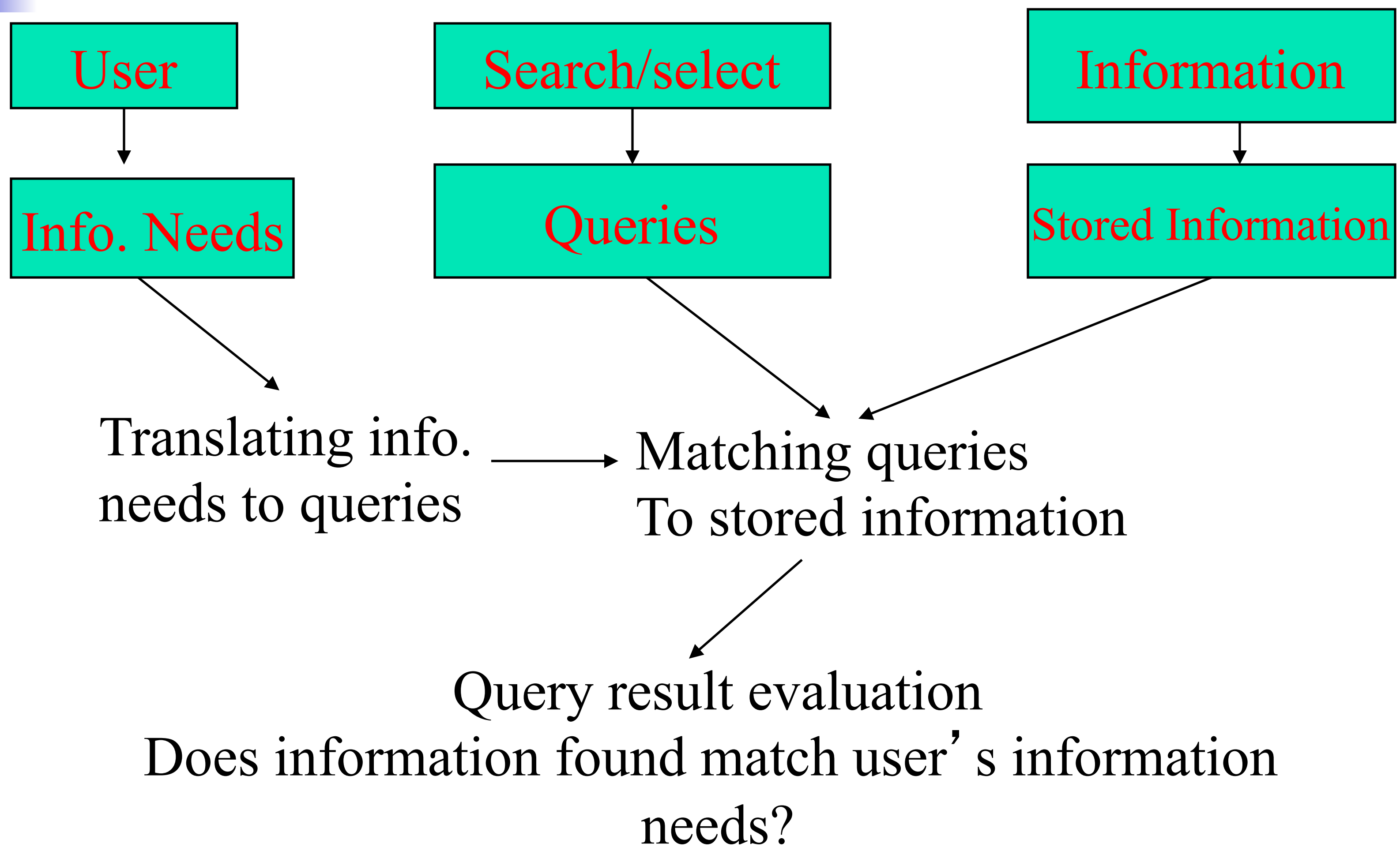
Text mining

- It refers to data mining using text documents as data.
- There are many special techniques for pre-processing text documents to make them suitable for mining.
- Most of these techniques are from the field of “Information Retrieval”.

Information Retrieval (IR)

- Conceptually, information retrieval (IR) is the study of finding needed information. i.e., IR helps users find information that matches their information needs.
- Historically, information retrieval is about document retrieval, emphasizing document as the basic unit.
- Technically, IR studies the acquisition, organization, storage, retrieval, and distribution of information.
- IR has become a center of focus in the Web era.

Information Retrieval



Text Processing

- Word (token) extraction
- Stop words
- Stemming
- Frequency counts

Stop words

- Many of the most frequently used words in English are worthless in IR and text mining – these words are called *stop words*.
  - the, of, and, to, ….
  - Typically about 400 to 500 such words
  - For an application, an additional domain specific stop words list may be constructed
- Why do we need to remove stop words?
  - Reduce indexing (or data) file size
    - stopwords accounts 20-30% of total word counts.
  - Improve efficiency
    - stop words are not useful for searching or text mining
    - stop words always have a large number of hits

Stemming

- Techniques used to find out the root/stem of a word:
  - E.g.,
    - user
    - users
    - used
    - using

    engineering
    engineered
    engineer
- stem: use engineer

Usefulness

- improving effectiveness of IR and text mining
  - matching similar words
- reducing indexing size
  - combing words with same roots may reduce indexing size as much as 40-50%.

Basic stemming methods

- remove ending
  - if a word ends with a consonant other than s,
    followed by an s, then delete s.
  - if a word ends in es, drop the s.
  - if a word ends in ing, delete the ing unless the remaining word consists only of one letter or of th.
  - If a word ends with ed, preceded by a consonant, delete the ed unless this leaves only a single letter.
  - ……
- transform words
  - if a word ends with "ies" but not "eies" or "aies" then "ies --> y."

Frequency counts

- Counts the number of times a word occurred in a document.
- Counts the number of documents in a collection that contains a word.
- Using occurrence frequencies to indicate relative importance of a word in a document.
  - if a word appears often in a document, the document likely “deals with” subjects related to the word.

Vector Space Representation

- A document is represented as a vector:
  - $(W_1, W_2, \ldots \ldots, W_n)$
- Binary:
  - $W_i =$ 1 if the corresponding term i (often a word) is in the document
  - $W_i =$ 0 if the term i is not in the document
- TF: (Term Frequency)
  - $W_i = tf_i$ where tf_i is the number of times the term occurred in the document
- TF*IDF: (Inverse Document Frequency)
  - $W_i = tf_i * idf_i = tf_i * \log(N/df_i))$ where df_i is the number of documents contains term i, and N the total number of documents in the collection.

Vector Space and Document Similarity

- Each indexing term is a dimension. A indexing term is normally a word.
- Each document is a vector
  - $D_i = (t_{i1}, \; t_{i2}, \; t_{i3}, \; t_{i4}, \; \ldots \; t_{in})$
  - $D_j = (t_{j1}, \; t_{j2}, \; t_{j3}, \; t_{j4}, \ldots, \; t_{jn})$
- Document similarity is defined as

$$\text{Similarity}\;(D_i, D_j) = \frac{\sum_{k=1}^{n} t_{ik} * t_{jk}}{\sqrt{\sum_{k=1}^{n} t_{ik}^{\;2}} \times \sqrt{\sum_{k=1}^{n} t_{jk}^{\;2}}}$$

Query formats

- Query is a representation of the user’s information needs
  - Normally a list of words.
- Query as a simple question in natural language
  - The system translates the question into executable queries
- Query as a document
  - “Find similar documents like this one”
  - The system defines what the similarity is

An Example

- A document Space is defined by three terms:
  - hardware, software, users
- A set of documents are defined as:
  - A1=(1, 0, 0), A2=(0, 1, 0), A3=(0, 0, 1)
  - A4=(1, 1, 0), A5=(1, 0, 1), A6=(0, 1, 1)
  - A7=(1, 1, 1) A8=(1, 0, 1). A9=(0, 1, 1)
- If the Query is “hardware and software”
- what documents should be retrieved?

An Example (cont.)

- In Boolean query matching:
  - document A4, A7 will be retrieved (“AND”)
  - retrieved:A1, A2, A4, A5, A6, A7, A8, A9 (“OR”)
- In similarity matching (cosine):
  - q=(1, 1, 0)
  - S(q, A1)=0.71, S(q, A2)=0.71, S(q, A3)=0
  - S(q, A4)=1, S(q, A5)=0.5, S(q, A6)=0.5
  - S(q, A7)=0.82, S(q, A8)=0.5, S(q, A9)=0.5
  - Document retrieved set (with ranking)=
    - {A4, A7, A1, A2, A5, A6, A8, A9}

Cosine Similarity

- Cosine Similarity is a technique that is derived from vector theory.
- In Information Retrieval, it is used to indicate (or measure)
  - the degree of similarity between two documents, or
  - between a document and a query.

- Keywords (Terms)
  - to describe the information content within a document.
- Vocabulary (Dictionary)
  - The total set of keywords
- Stopwords
  - Words which do not help to differentiate documents or which don’t identify the information within a document
  - Discarded from the list of keywords

- Examp

$$
\begin{pmatrix} Car \\ Van \\ Bus \\ Road \\ Highway \\ Bicycle \\ Coach \\ Train \\ Station \\ Ticket \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 1 \\ 0 \\ 1 \\ 2 \\ 1 \\ 1 \end{pmatrix}
$$

- Vector Inner Product (Dot Product)
  - is defined as the *sum of the products* of the vector components.

$$\text{let}\quad \mathbf{v} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix} \quad \text{and} \quad \mathbf{w} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

$$\mathbf{v}.\mathbf{w} = (3.1) + (-1.2) + (2.1) = 3$$

- Vector Length (Norm)
  - Inner product of vector with itself

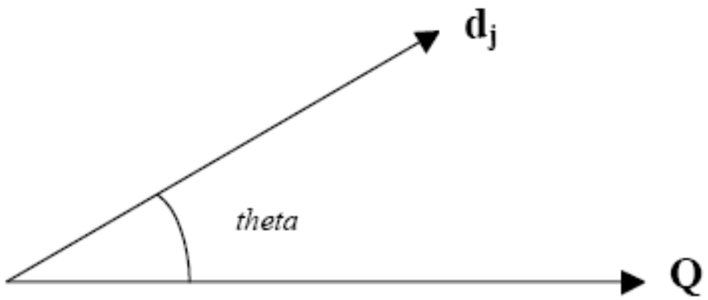
$$||\mathbf{v}|| = (\mathbf{v}.\mathbf{v})^{1/2} \text{ or } \sqrt{(\mathbf{v}.\mathbf{v})}$$

$$\mathbf{v} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

$$\mathbf{v}.\mathbf{v} = 3^2 + 4^2 = 25$$

  - Take the square root
    - Length = 5

- Cosine Similarity



$$\cos\theta = \frac{\mathbf{v}.\mathbf{w}}{||\mathbf{v}||.||\mathbf{w}||}$$

$$\cos\theta = \frac{\text{inner product of vectors } \mathbf{v},\mathbf{w}}{(\text{length of vector } \mathbf{v}).(\text{length of vector } \mathbf{w})}$$

- Example

$$\mathbf{v} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad \mathbf{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\mathbf{v}.\mathbf{w} = (0.1) + (1.1)$$

$$||\mathbf{v}|| = \sqrt{(0^2 + 1^2} = 1$$

$$||\mathbf{w}|| = \sqrt{(1^2 + 1^2} = \sqrt{2}$$

$$\cos\theta = \frac{1}{1.\sqrt{2}} = 0.707$$

$$\theta = \cos^{-1}(0.707) = 45^\circ$$

For identical documents

- v = (1, 1) and w = (1, 1)
- v.w = (1.1) + (1.1) = 2
- $\| v \| = \sqrt{2}$
- $\| w \| = \sqrt{2}$
- $\cos \theta = \dfrac{2}{\sqrt{2}\sqrt{2}} = 1$
- $\theta = 0°$

For dissimilar documents

- v = (1, 0) and w = (0, 1)
- v.w = (1.0) + (0.1) = 0
- ║v║ = 1
- ║w║ = 1

- $\cos\theta = \dfrac{0}{\sqrt{1}\sqrt{1}} = 0$

- $\theta = 90^\circ$

- For *n*-dimensional vectors

$$\cos\theta = \frac{\sum_{i=1}^{n} \mathbf{v}_i . \mathbf{w}_i}{\left(\sum_{i=1}^{n} (\mathbf{v}_i)^2\right)^{1/2} \left(\sum_{i=1}^{n} (\mathbf{w}_i)^2\right)^{1/2}}$$

- Example:
  - 3-dimensions
  - $v = (v_1, v_2, v_3)$ and $w = (w_1, w_2, w_3)$

$$\cos\theta = \frac{(\mathbf{v}_1 . \mathbf{w}_1) + (\mathbf{v}_2 . \mathbf{w}_2) + (\mathbf{v}_3 . \mathbf{w}_3)}{\sqrt{[(\mathbf{v}_1)^2 + (\mathbf{v}_2)^2 + (\mathbf{v}_3)^2]} . \sqrt{[(\mathbf{w}_1)^2 + (\mathbf{w}_2)^2 + (\mathbf{w}_3)^2]}}$$

$$\mathbf{d}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \mathbf{d}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \mathbf{d}_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

- Example

$$\cos\theta_{12} = \frac{(\mathbf{d}_{11}.\mathbf{d}_{21}) + (\mathbf{d}_{12}.\mathbf{d}_{22}) + (\mathbf{d}_{13}.\mathbf{d}_{23})}{\sqrt{[(\mathbf{d}_{11})^2 + (\mathbf{d}_{12})^2 + (\mathbf{d}_{13})^2]}\,.\,\sqrt{[(\mathbf{d}_{21})^2 + (\mathbf{d}_{22})^2 + (\mathbf{d}_{23})^2]}}$$

$$= \frac{(1.1) + (0.1) + (1.1)}{\sqrt{[(1)^2 + (0)^2 + (1)^2]}\,.\,\sqrt{[(1)^2 + (1)^2 + (1)^2]}}$$

$$\cos\theta_{12} = \frac{1+0+1}{\sqrt{2}\,.\,\sqrt{3}} = 0.82 \quad (35°)$$

d1*d2	1	1	2	**d1*d3**	1	0	1	**d2*d3**	1	0	2
	0	1	2.449		0	1	2		1	1	2.449
	1	1			1	1			1	1	
cosθ =			**0.82**				**0.5**				**0.82**
angle =			**35**				**60**				**35**

Using one or more queries

$$\mathbf{q}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \mathbf{q}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \mathbf{q}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

d1*q1	1	0	1	**d2*q1**	1	0	1	**d3*q1**	0	0	1
	0	0	1.414		1	0	1.732		1	0	1.414
	1	1			1	1			1	1	
cosθ =			**0.707**				**0.577**				**0.707**
angle =			**45**				**55**				**45**

d1*q2	1	0	0	**d2*q2**	1	0	1	**d3*q2**	0	0	1
	0	1	1.414		1	1	1.732		1	1	1.414
	1	0			1	0			1	0	
cosθ =			**0**				**0.577**				**0.707**
angle =			**90**				**55**				**45**

d1*q3	1	1	1	**d2*q3**	1	1	1	**d3*q3**	1	1	1
	0	0	1.414		1	0	1.732		1	0	1.414
	1	0			1	0			0	0	
cosθ =			**0.707**				**0.577**				**0.707**
angle =			**45**				**55**				**45**

- Example

$$\mathbf{d}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad \mathbf{d}_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad \mathbf{d}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\mathbf{q}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{q}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{q}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

d1*q1	1	1	1	d2*q1	0	1	0	d3*q1	1	1	1
	1	0	2.828		0	0	2.236		1	0	3.162
	1	0			0	0			1	0	
	0	0			1	0			1	0	
	1	0	0.354		0	0	0		1	0	0.316
	0	0			0	0			1	0	
	1	0			1	0			1	0	
	1	0			1	0			1	0	
	1	0			1	0			1	0	
	1	0			1	0			1	0	

d1*q2	1	1	3	d2*q2	0	1	0	d3*q2	1	1	3
	1	0	4.899		0	0	3.873		1	0	5.477
	1	1			0	1			1	1	
	0	0			1	0			1	0	
	1	1	0.612		0	1	0		1	1	0.548
	0	0			0	0			1	0	
	1	0			1	0			1	0	
	1	0			1	0			1	0	
	1	0			1	0			1	0	
	1	0			1	0			1	0	

d1*q3	1	1	4	d2*q3	0	1	3	d3*q3	1	1	5
	1	0	6.325		0	0	5		1	0	7.071
	1	0			0	0			1	0	
	0	0			1	0			1	0	
	1	0	0.632		0	0	0.6		1	0	0.707
	0	1			0	1			1	1	
	1	1			1	1			1	1	
	1	0			1	0			1	0	
	1	1			1	1			1	1	
	1	1			1	1			1	1	

$$\mathbf{d}_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 1 \\ 0 \\ 1 \\ 2 \\ 1 \\ 1 \end{pmatrix} \quad \mathbf{d}_2 = \begin{pmatrix} 0 \\ 4 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 2 \\ 1 \\ 1 \end{pmatrix} \quad \mathbf{d}_3 = \begin{pmatrix} 2 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{pmatrix}$$

$$\mathbf{q}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{q}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{q}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

d1*q1	1	1	1	**d2*q1**	0	1	0	**d3*q1**	2	1	2
	2	0	4.796		4	0	4.899		1	0	4.123
	1	0			0	0			1	0	
	3	0			1	0			0	0	
	1	0	**0.209**		0	0	**0**		1	0	**0.485**
	0	0			0	0			1	0	
	1	0			1	0			2	0	
	2	0			2	0			0	0	
	1	0			1	0			1	0	
	1	0			1	0			2	0	

d1*q2	1	1	3	**d2*q2**	0	1	0	**d3*q2**	2	1	4
	2	0	8.307		4	0	8.485		1	0	7.141
	1	1			0	1			1	1	
	3	0			1	0			0	0	
	1	1	**0.361**		0	1	**0**		1	1	**0.56**
	0	0			0	0			1	0	
	1	0			1	0			2	0	
	2	0			2	0			0	0	
	1	0			1	0			1	0	
	1	0			1	0			2	0	

d1*q3	1	1	4	**d2*q3**	0	1	3	**d3*q3**	1	1	6
	2	0	10.724		4	0	10.954		1	0	8.367
	1	0			0	0			1	0	
	3	0			1	0			0	0	
	1	0	**0.373**		0	0	**0.274**		2	0	**0.717**
	0	1			0	1			1	1	
	1	1			1	1			1	1	
	2	0			2	0			0	0	
	1	1			1	1			1	1	
	1	1			1	1			2	1	

Relevance judgment for IR

- A measurement of the outcome of a search or retrieval
- The judgment on what should or should not be retrieved.
- There is no simple answer to what is relevant and what is not relevant: need human users.
  - difficult to define
  - subjective
  - depending on knowledge, needs, time,, etc.
- The central concept of information retrieval

Precision and Recall

- Given a query:
  - Are all retrieved documents relevant?
  - Have all the relevant documents been retrieved ?
- Measures for system performance:
  - The first question is about the precision of the search
  - The second is about the completeness (recall) of the search.

Precision and Recall (cont)

	Relevant	Not Relevant
Retrieved	a	b
Not retrieved	c	d

$$P = \frac{a}{a+b}$$

$$R = \frac{a}{a+c}$$

Precision and Recall (cont)

- Precision measures how precise a search is.
  - the higher the precision,
  - the less unwanted documents.

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

- Recall measures how complete a search is.
  - the higher the recall,
  - the less missing documents.

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of all the relevant documents in the database}}$$

Relationship of R and P

- Theoretically,
  - R and P not depend on each other.
- Practically,
  - High Recall is achieved at the expense of precision.
  - High Precision is achieved at the expense of recall.
- When will p = 0?
  - Only when none of the retrieved documents is relevant.
- When will p=1?
  - Only when every retrieved documents are relevant.
- Depending on application, you may want a higher precision or a higher recall.

Alternative measures

- Combining recall and precision, F score

$$F = \frac{2PR}{R + P}$$

- Breakeven point: when p = r
- These two measures are commonly used in text mining: classification and clustering.
- Accuracy is not normally used in text domain because the set of relevant documents is almost always very small compared to the set of irrelevant documents.

Web Search as a huge IR system

- A Web crawler (robot) crawls the Web to collect all the pages.
- Servers establish a huge inverted indexing database and other indexing databases
- At query (search) time, search engines conduct different types of vector query matching

Different search engines

- The real differences among different search engines are
  - their indexing weight schemes
  - their query process methods
  - their ranking algorithms
  - None of these are published by any of the search engines firms.