Information Retrieval and Data Mining

Prof. Marco Tagliasacchi
Prof. Matteo Matteucci

February, 25 2015

1. **Question (8 pts)**: Consider a graph represented by the following adjacency matrix

$$\mathbf{E} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \tag{1}$$

   (a) Compute the PageRank score associated to each node in the graph, setting $\mu = 0$. You are asked to compute the exact solution, not an approximation.

   (b) Can you guarantee that PageRank converges to a unique solution by looking at the matrix E? Does PageRank converge to a unique solution in this case (once scores are normalized so that they sum up to one)?

   (c) Compute the probability that a random surfer is visiting either node 1 or node 3 at some future time $t \to \infty$.

   (d) Compute the PageRank score associated to each node in the graph, setting $\mu = 0.5$. Does this affect the probability that you computed above?

2. **Question (6 pts)**

   (a) Describe Kemeny's rank aggregation.

   (b) Discuss the computational complexity of the aggregation algorithm using either Kandall tau or footrule distance.

   (c) Describe median rank aggregation, explaining when this is an approximation of Kemeny's rank aggregation.

3. **Questions (5 pts - each statement can be either TRUE or FALSE)**

(a) Consider a document collection that is represented by means of the following term-document matrix

$$A = \begin{bmatrix} 4 & 8 & 2 & 4 \\ 0 & 1 & 2 & 0 \\ 1 & 0 & 0 & 2 \\ 8 & 2 & 8 & 1 \\ 4 & 0 & 0 & 0 \end{bmatrix} \tag{2}$$

Let $\mathbf{q} = [1, 0, 0, 0, 1]^T$ denote the query vector.

- T F If $w_{ij} = freq_{ij}$, then the cosine similarity between document d_1 and q is equal to $8/\sqrt{192}$
- T F When idf is adopted, then $w_{4,j} = w_{1,j}$, $\forall j$.
- T F Considering a Boolean model, the relevant set is $\{d_1\}$
- T F When the query and the documents are normalized so that they sum up to one, ranking by decreasing cosine similarity is the same as ranking by increasing Euclidean distance.

(b) Consider a dataset of N vectors in a d-dimensional space and a query vector $\mathbf{q}$.

- T F Finding the nearest neighbor to $\mathbf{q}$ requires visiting on average at least $O(N^{1/2})$ points.
- T F When $d \to \infty$, it can happen that finding the nearest neighbor requires visiting $O(N)$ points.
- T F In a k-d tree, each node of the tree has exactly 2^d children.

(c) Consider two queries that produce the following rankings: $< a, b, c, d, e >$, $< c, e, f, l, b >$. The ground-truth relevant sets are $R_{q_1} = \{a, c\}$, $R_{q_2} = \{c, e, l\}$

- T F $MAP = \frac{21}{24}$.
- T F Precision at 2 is equal to 1/2.
- T F In this case, precision is always decreasing with k

4. **Question (8 pts)** Answer the following questions:

1. What is the goal of frequent pattern mining (a.k.a. association rules mining)? (1 point)
2. How Support and Confidence are defined? (1 point)
3. Describe the a-priori algorithm for *Frequent Itemset* and *Rule Generation*. (2 points)
4. Starting from the table find all the itemset with minimum support 60%. (2 points)
5. Out of the biggest itemsets extract the association rules with 100% confidence. (2 points)

TID	A	B	C	D	E
T1	1	1	1	0	0
T2	1	1	1	1	1
T3	1	0	1	1	0
T4	1	0	1	1	1
T5	1	1	1	1	0

5. **Question (5 pts)** Lets make some effort in making sense of what we have learned so far in Data Mining! With reference to the picture explain in details one real-life example in which you expect to apply the topics learned in the course. Use your imagination, we appreciate if you try to make this example as close as possible to the job you would like to do after graduation!

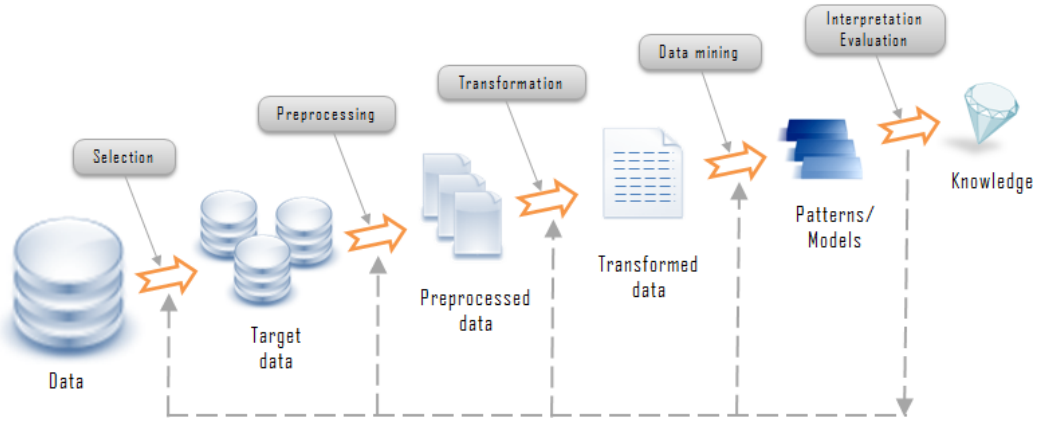

Figure 1: The Knowledge Discovery Process

Please note that we are not asking to describe the picture, but to describe clearly and in details an example for which each of the previous (5) steps could been applied and how.