Information Retrieval and Data Mining

Prof. Marco Tagliasacchi
Prof. Matteo Matteucci

September, 29 2015

Very Important Notes

- Answers to questions 1, 2, and 3 should be delivered on a different sheet with respect to 4 and 5
- If you need a calculator this should not be to any extent programmable or network connected

1. **Question (8 pts)**: Consider a document collection represented by the following term-document matrix, where W_{ij} represents the td-idf relevance score of term i in document j.

$$W = \begin{bmatrix} 1 & 2 & 4 & 4 & 1 & 0 \\ 1 & 0 & 0 & 0 & 2 & 2 \\ 0 & 1 & 3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 \end{bmatrix} \tag{1}$$

   Assume that LSI was performed, leading to the following term-topic matrix:

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2}$$

   Assume $\sigma_1 = \sigma_2 = 1$.

   (a) Consider the query $q = [1010]$. Compute the cosine similarity between the query and each document, and produce the ranked result set.

   (b) Draw a 2d chart showing the position of the documents in the topic space.

   (c) Consider the query $q = [1010]$. Plot the query in the topic space and show geometrically (this time without explicitly doing any calculation!) what is the the ranked result set according to cosine similarity.

2. **Question (6 pts)**

   Create a graph with 4 nodes. Compute the corresponding adjacency matrix and the PageRank score of each node (assume damping factor $\mu = 0.1$).

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

(a) Consider the following rankings (break the tie in favour of the object with the smallest index).

k	rank 1	rank 2	rank 3
1	a	a	b
2	b	c	a
3	c	d	c
4	d	e	f
5	e	b	e
6	f	f	d

- [T] [F] Object a is the Condorcet's winner.
- [T] [F] $d_F(r_1, r_2) < d_F(r_2, r_3)$, where $d_F(\cdot, \cdot)$ denotes the footrule distance between two rankings.
- [T] [F] The median rank aggregation is $< a, b, c, d, e, f >$.
- [T] [F] $d_K(r_1, r_2) = 1/5$

(b) Consider an inverted index.

- [T] [F] Each posting list includes documents containing a given term.
- [T] [F] Stemming reduces the number of terms to be stored in the dictionary.
- [T] [F] Accessing the dictionary requires $O(log(M))$ operations, where M is the number of terms.
- [T] [F] The dictionary is preferably stored in main memory.

(c) Consider the following rankings induced by the scores, and accessing to the data by means of the threshold algorithm (TA) to find the top-1 object. Let the aggregation function be the arithmetic average of the scores $s = (s_1 + s_2)/2$.

obj	s_1	obj	s_2
a	0.98	b	0.90
b	0.92	a	0.88
c	0.70	c	0.86

- [T] [F] After retrieving 1 elements from ranking 1, and 2 element from ranking 1, the threshold is equal to 0.93.
- [T] [F] The threshold algorithm stops after retrieving 2 elements from ranking 1, and 1 element from ranking 2.
- [T] [F] By changing the aggregation function, the aggregated ranking remains the same.

4. **Question (8 pts)**

Imagine we have a data set where each record is a list of categorical weather conditions on a randomly selected number of days, and the labels correspond to whether a girl named Arya went for a horse ride on that day.

(a) **Describe in details** the Decision Tree model and its training algorithm [2 points]

Sky	Temperature	Humidity	Wind	Horse Ride
Cloudy	Warm	Low	Low	Yes
Rainy	Cold	Medium	Low	No
Sunny	Warm	Medium	Low	Yes
Sunny	Hot	High	High	No
Snow	Cold	Low	High	No
Rainy	Warm	High	Low	Yes

(b) **Learn a decision tree** out of the previous dataset by using *Information Gain*. [4 points]

(c) **Extract a Ruleset** out of the decision tree built at the previous step and describe how to prune it eliminating antecedents by means of the Chi-Square test for independence (no need to do the pruning, just explain how). [2 points]

5. **Question (5 pts)** Let consider the Knowledge Discovery Process, which starts from **data** to generate **knowledge**, reported in Figure 1

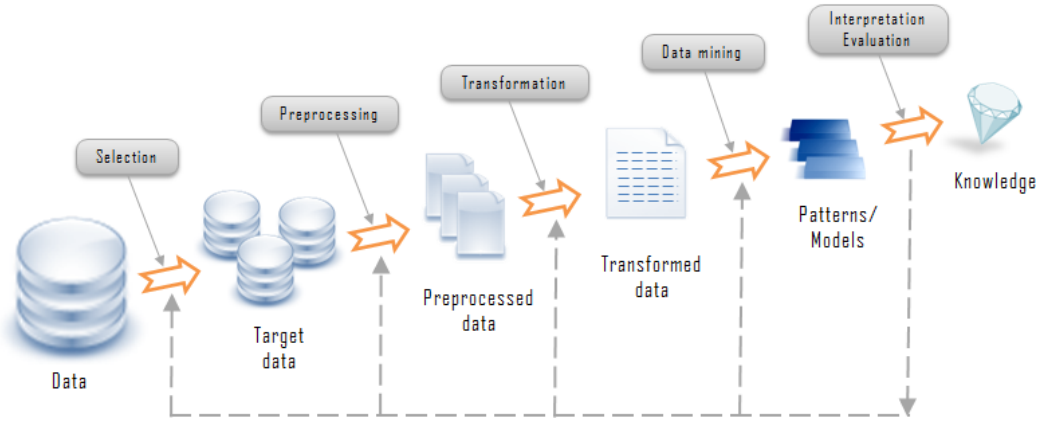

Figure 1: The Knowledge Discovery Process

(a) Discuss the 5 steps from `data` to `knowledge` in the figure and the reason for the corresponding feedback arrow [3 points]

(b) The outcome of this process is a pattern or a set of patterns. A pattern, to be good, should be `novel`, `useful`, and `understandable`; can you explain each of the previous terms and discuss how each of them is related/affected to/by the previous steps? (e.g., is the usefulness of a pattern mainly affected by the transformation step or the selection step?) [2 points]