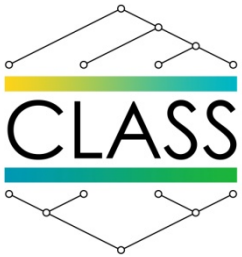


Co-funded by the
Erasmus+ Programme
of the European Union

INFORMATION RETRIEVAL

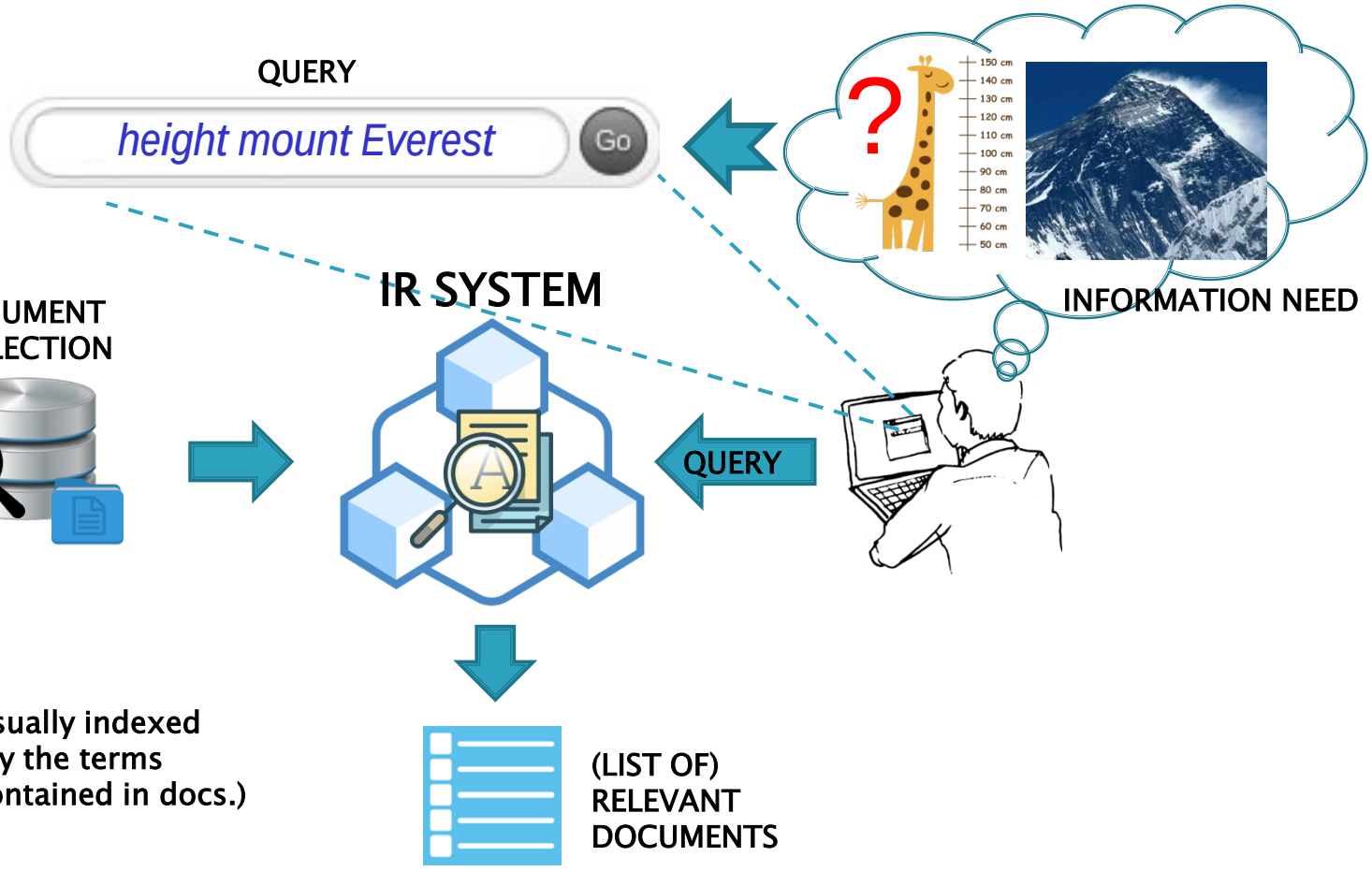
Prof. Jesús Vilares
jesus.vilares@udc.es

UNIVERSIDADE DA CORUÑA

Purpose

- Given a **document collection** and an **information need** of the user expressed as a text **query**, an **Information Retrieval (IR)** system will retrieve a [ranked] list of those documents which deal with the topic of the query (i.e. whose content potentially satisfies that original need).
  - It doesn't obtain the information required by the user, it just indicates where such information should be.
- **Applications:**
  - Document search services (e.g. libraries)
  - Internet Search Engines
  - Spam filters
  - (…)

Purpose


QUERY
height mount Everest
Go
INFORMATION NEED
IR SYSTEM
DOCUMENT COLLECTION
QUERY
Hockey 128
Hockey stick 128
Holly 118
Home plate 124
Homebase 124
Honor roll 139
Hoof 115
Hook 132
Hook/ladder 120
Horizontal bar 128
Hornet 114
Horoscope 139
Horse 115, 144
Horse Whip 115
Horseshoe 115
Usually indexed (by the terms contained in docs.)
(LIST OF) RELEVANT DOCUMENTS

The *Bag-of-Terms* paradigm

- Documents and queries are represented as **sets of *terms*** (a.k.a. *index terms* or *keywords*)
- If a text contains a given term, we can assume that, somehow, that text addresses such topic.
- If a **query and a document share** one or more **index terms**, we can assume that the document addresses the topic of the query.

Everest

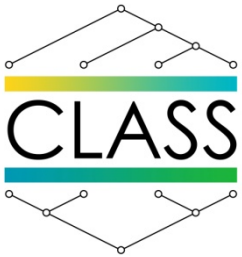
Everest

Go

Term weighting

- **Not all terms are equally representative or important**
- **Weight** w_{ij} of an index term t_i in document d_j
- Based on:
  - tf_{ij} : no. of occurrences of term t_i in document d_j
  - n_i : no. of documents of the collection containing term t_i
  - N : no. of documents of the collection
  - l_j : length of document d_j
- Diverse ways of combining them: **weighting schemes**
- Most popular (baseline): ***tf–idf***

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_i}$$



Retrieval models

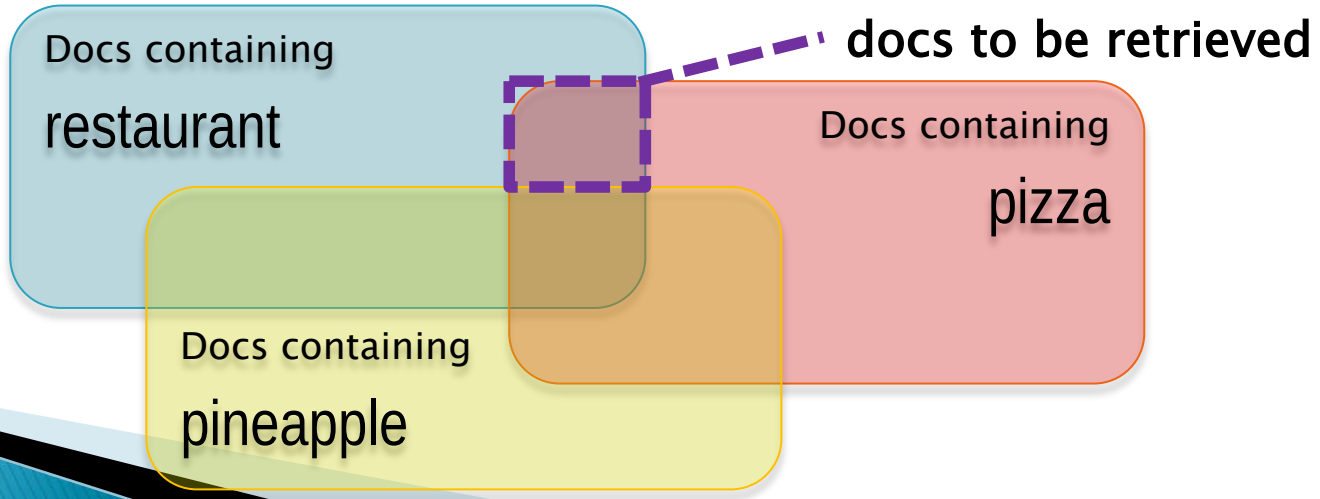
- A **retrieval model** establishes:
  1. How to represent a document
  2. How to represent an information need (query)
  3. How to compare them

- Classical models:
  - Boolean model
  - Vector model
  - Probabilistic model

Boolean model

- Mathematical basis: **Boolean algebra** + **Set theory**
- Document: **set of terms** contained in the text
- Query: **condition** expressed as **boolean expression** = terms + boolean operators (AND, OR, NOT)
- Docs. fulfilling the condition **totally** are retrieved with no ranking

e.g. restaurant AND pizza AND NOT pineapple

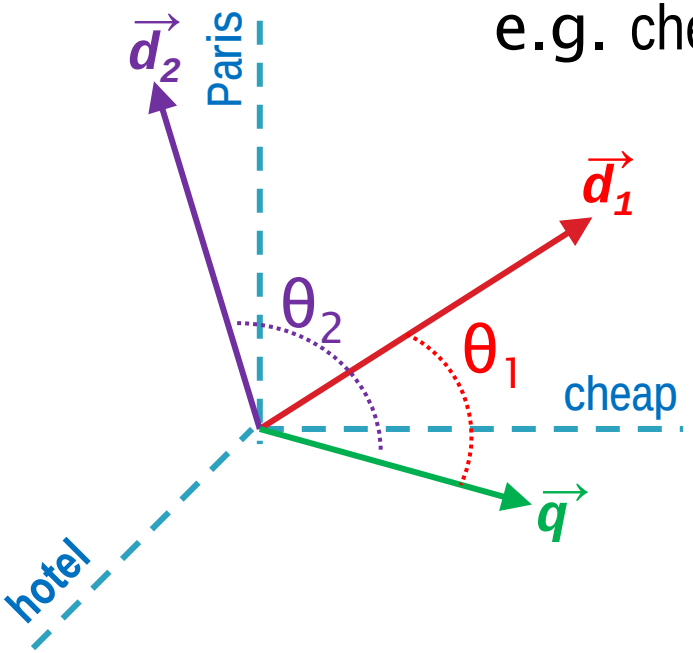


Vector model

- Mathematical basis: **linear algebra**
- Documents/queries: **vectors** in T–dimensional space
  - $T =$ size of vocabulary (i.e. number of unique terms)
  - Dimension i corresponds to term t_i
  - Doc d_j as vector $\overrightarrow{d_j} = (w_{1j}, w_{2j}, \cdots, w_{Tj})$
  - Query q as vector $\vec{q} = (w_{1q}, w_{2q}, \cdots, w_{Tq})$
  - w_{ij} : weight of index term t_i in document d_j

Vector model

e.g. cheap hotel Paris



- The closer the vector of document d_j is to the vector of query q, the more similar they are (i.e. the more relevant the document is)
  - Distance between vectors measured using the **cosine of the angle θ** formed by them

Vector model

- Distance between two vectors measured using the **cosine of the angle θ** formed by them:

$$\mathrm{sim}(d_j, q) = \cos\theta = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{T} w_{ij} \times w_{iq}}{\sqrt[2]{\sum_{i=1}^{T} w_{ij}^2} \times \sqrt[2]{\sum_{i=1}^{T} w_{iq}^2}}$$

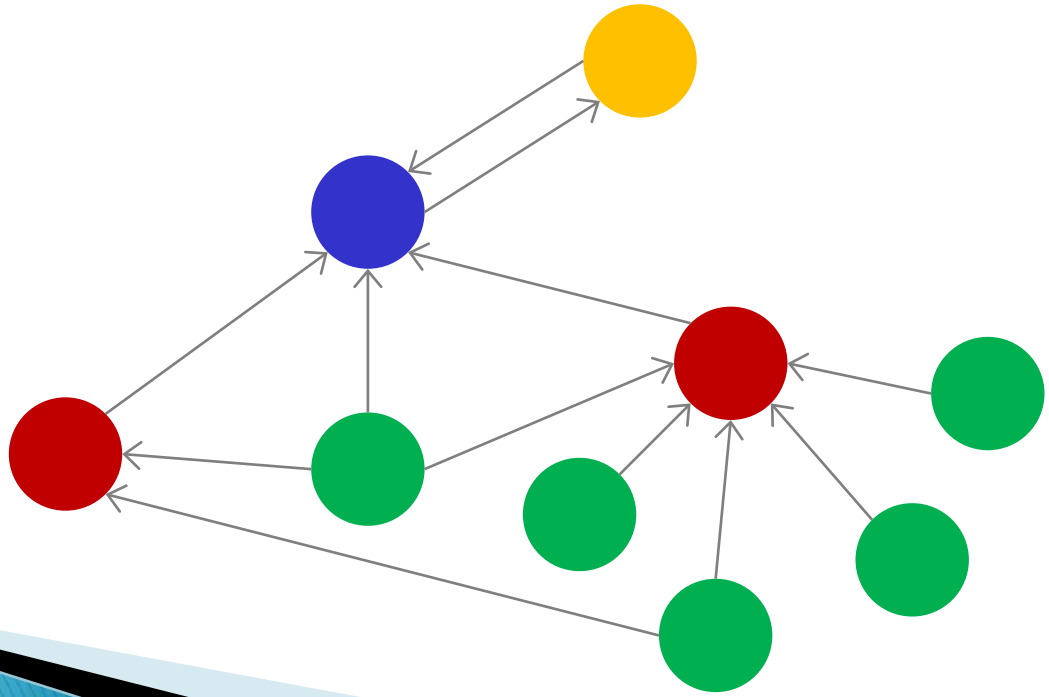
- The vector model allows:
  - Partial matching
  - Ranking of results (according to similarity)

Web IR

- In the Web, documents are **connected through hyperlinks** forming a huge graph.
- Classical IR models do not consider this.
- A good Web IR model must also take into account:
  - Structure of web pages
  - *Anchor texts* of the links (they describe the target page)
  - **Popularity** of each web page
  - (…)
- The so-named **PageRank algorithm** developed by Google was a major technological breakthrough.

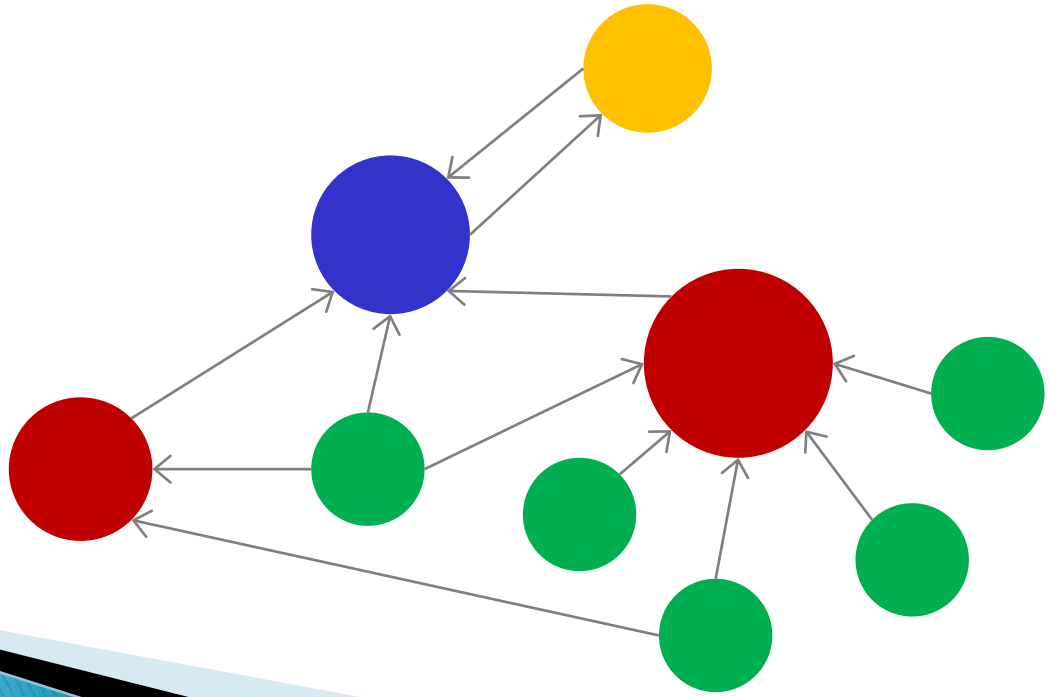
PageRank

- The popularity of a web page is computed according to the number of incoming links received



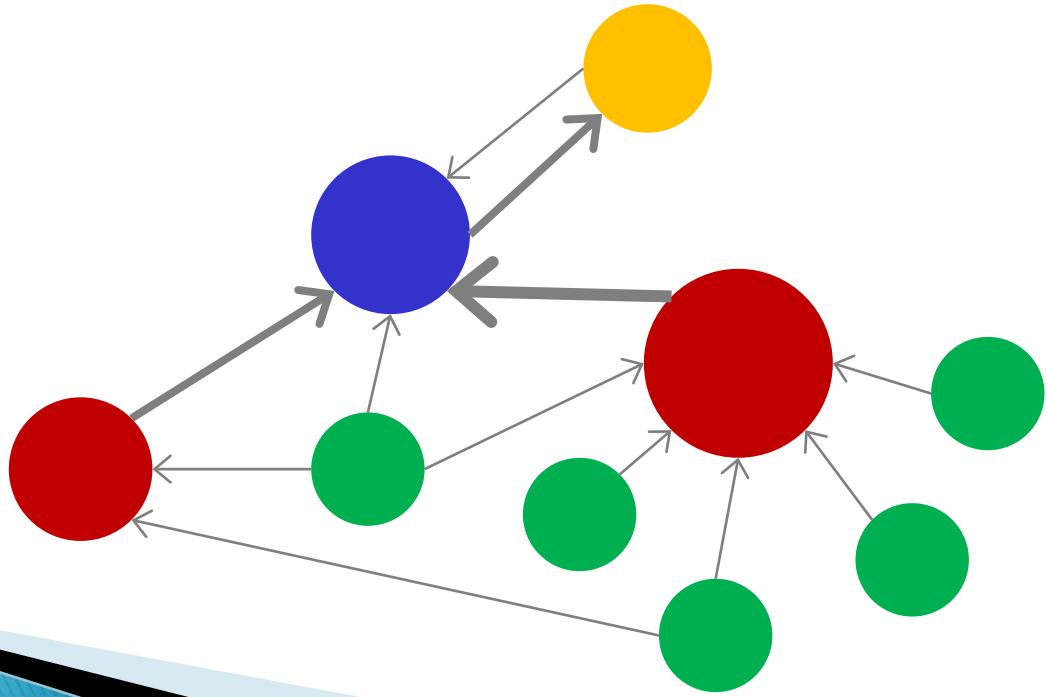
PageRank

- A **link from page A to page B** can be seen as **a vote** given by A to B
- **The more votes** (incoming links) you have, **the more popular** you are



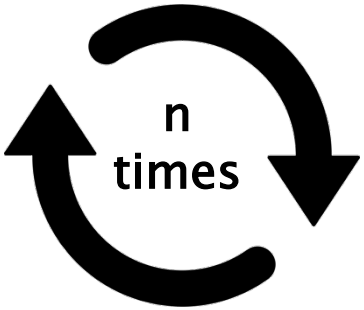
PageRank

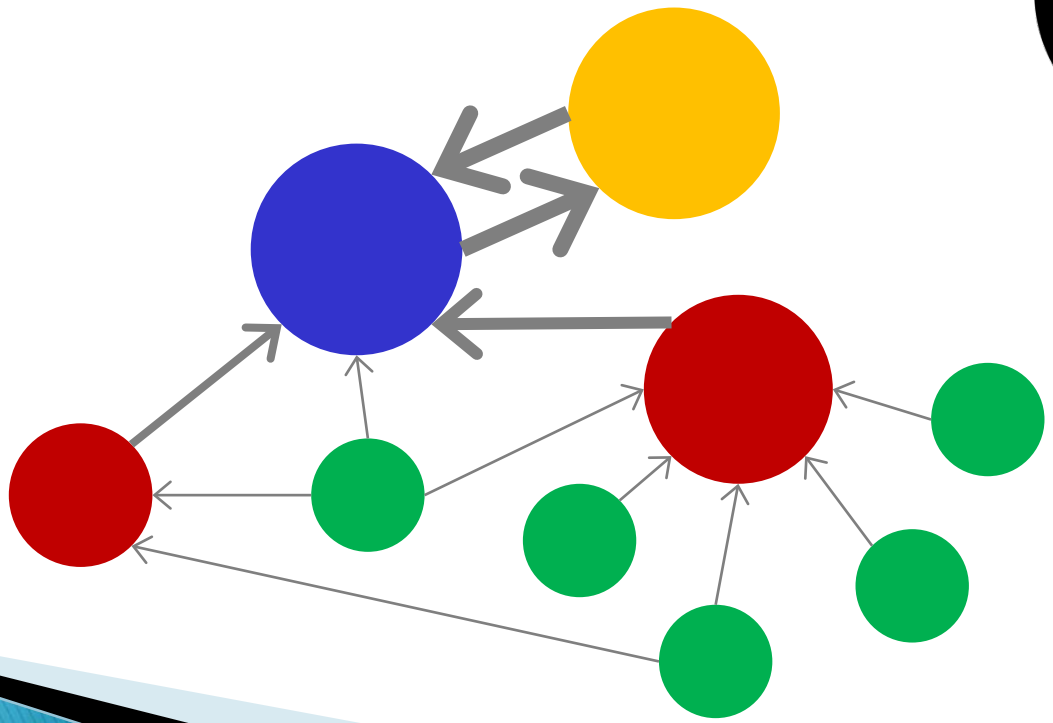
- Not all votes have the same value
- A vote **(incoming link) from a popular page is more valuable**



PageRank

- Feedback is present; thus, PageRank must be calculated iteratively





Bibliography

- **[Baeza-Yates & Ribeiro-Neto, 2011]** Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: the concepts and technology behind search (2nd edition)*. Pearson Education.
- **[Büttcher et al., 2010]** Büttcher, S., Clarke, C.L.A. & Cormack, G.V. (2010). *Information Retrieval. Implementing and Evaluating Search Engines*. MIT Press.
- **[Croft et al., 2009]** Croft, W.B., Metzler, D. & Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Pearson Education. Freely available at: https://ciir.cs.umass.edu/irbook/
- **[Manning et al., 2008]** Manning, C.D., Raghavan, P. & Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge University Press.