Líneas de investigación en web mining, extracción automática de conocimiento y redes sociales

Carlos García Figuerola, José Luis Alonso Berrocal, Ángel Zazo Rodríguez

Instituto Universitario de Estudios de la Ciencia y la Tecnología, Universidad de Salamanca

{figue, berrocal, angelzazo}@usal.es

Appliying Social Network Analysis to Topic Detection & Tracking

Carlos G. Figuerola - ECyT Institute
University of Salamanca
II Seminario E-lectra - 25/04/2013

Topic Detection & Tracking

- Techniques for detecting the appearance of new topics and for tracking the reapparance and evolution of them.
- The typical scenario is a broadcast news system delivering speech and text in real time.
- These techniques can be applied in another similar fields: twitter and other social networks, digital archives, digital news press ...

Topic Detection & Trackinkg

Involves several techniques:

- speech to text conversion
- text's segmentation
- text's clustering
- detection of seminal stories
- tracking temporal evolution of topics

Clustering of text documents

- It's main target is to group similar documents together
- Documents (tweets, news stories, ..) talking about a same topic should be clustered together

However:

- clustering of documents is computationally expensive
- most of *clasic algorithms* (k-means, etc.) require establish the number of desired clusters in advance
- they have problems with noise (ie: ambiguous documents, marginal or minoritary topics)

Social Networks Analysis Techniques

- objects can establish relationships between them
- we can map objects and relationships towards a network or graph
  - objets are nodes
  - relationships are edges or links between nodes

Social Networks Analysis Techniques

Graph Theory can by applied to a network. We have tools to:

- evaluate features of individual nodes
- characterize the whole network
- to find out communities of nodes

Communities in Networks

In a network, a community is a bunch of nodes strongly linked between them and weakly with the others

There are several algorithms to find out communities. Differences lie in:

- speed
- take in account direction and weights of links
- bias to produce many and little or few and bigger communities


Information Technologies
Paleonthology
Maths
Zoology/Bothanycal
Foods
Health Sci
Aerospace tech.
Astronomy
engineering
architecture
Chemistry
Physics
Geometry
Earth Sci

Detecting Topics in digital press about Sci & Tech

- work in progress !
- we collected about 1 M of news from several spanish digital newspapers (2002-2011)
- applying automatic categorization techniques, we were able to select 50 K news about S & T
- these are the spanish part of the MACAS project (Mapping the Cultural Authority of Science)

Representing news (text documents) as a Network

We can try to deal with our news on S & T as they were a graph or network.

- each of news stories is a node
- we can establish a link between a pair of news stories if they are semantically correlated
- links are *undirected*
- links have *weigth*, reflecting the intensity of semantic similarity (as the semantic similarity is different for every pair of news)

Computing similarity between news

We use concepts and techniques borrowed from classic Information Retrieval discipline:

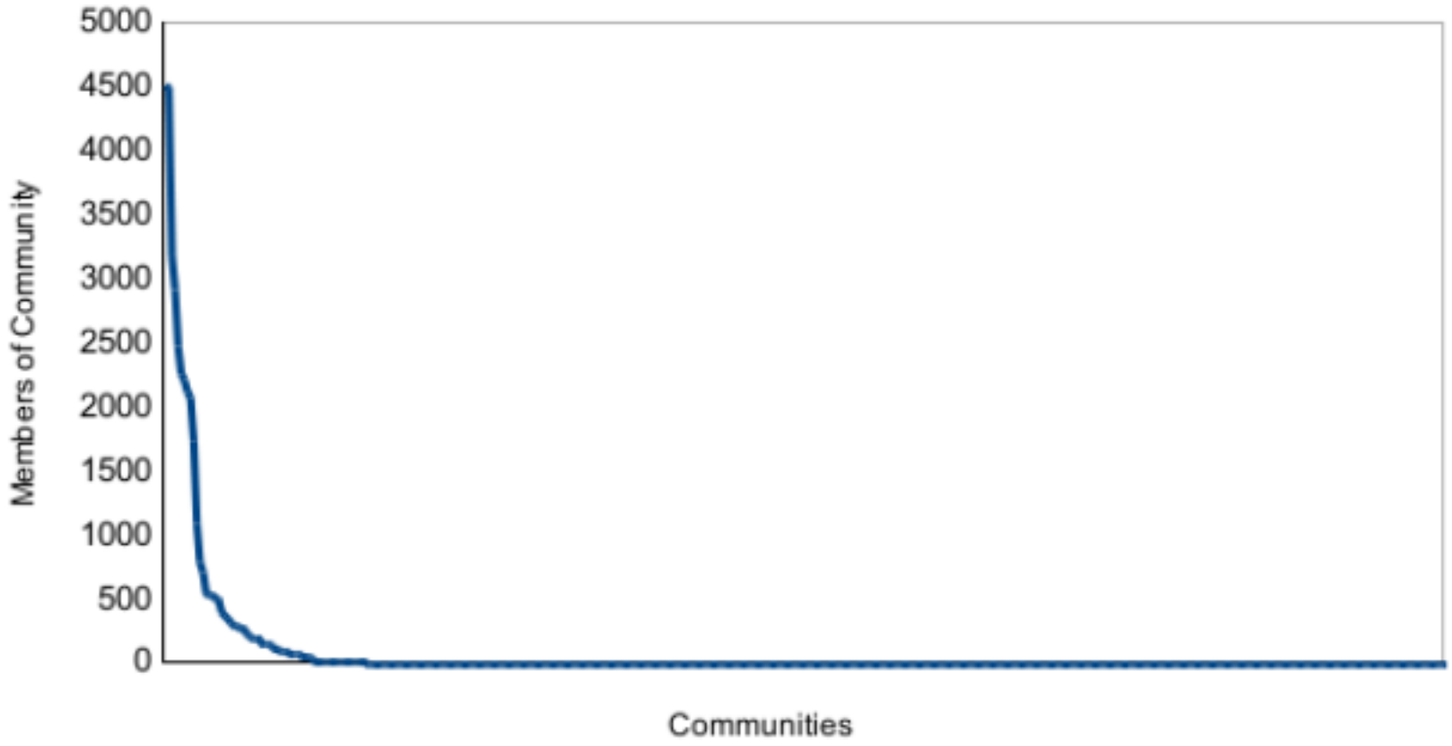
- from the vectorial model
  - each of news is a vector of terms
  - each term in each of news has weight
  - the similarity between a pair of news is the similairity between their vectors of terms
  - similarity is a value >= 0, <=1

The strength of a link between a pair of nodes (news) is their similarity

Detecting communities

After several tries, we choosen the Infomap algorithm to detect communities in our network

- it is very fast and memory efficient (we have 50 K nodes and more than 3 M of links!)
- it works fine with weights or strenght of links
- it produces an hierachical tree of communities, but only 3 sublevels
- often third level is too fine grained, so it is unuseful



in our case, InfoMap produced only 23 communities at the first level

Analyzing Results

Communities listing

community	topic
1	Public Health
2	Biomedicine
3	Energy
4	Human Development
5	Natural Resources
6	Aerospace Research
7	Biodiversity
8	Astronomy & Cosmology
9	Information Technology
10	Science Policy
11	Protected Species - Spain
12	Human Evolution
13	Contamination

Analyzing Results

Subcommunity	Topic	Subcommunity	Topic
1.1	influenza	1.11	infections, E. Coli,
1.2	AIDS	1.12	cholera
1.3	mortality	1.13	Legionella
1.4	drugs	1.14	polio
1.5	vaccines	1.15	mad cow disease
1.6	malaria	1.16	foot and mouth disease
1.7	SARS	1.17	dengue
1.8	tuberculosis	1.18	insect infections
1.9	hepatitis C	1.19	Chagas
1.10	antibiotics, bacteria	1.20	bio-bac