WIKICLUSTER

Labeling clusters with Wikipedia

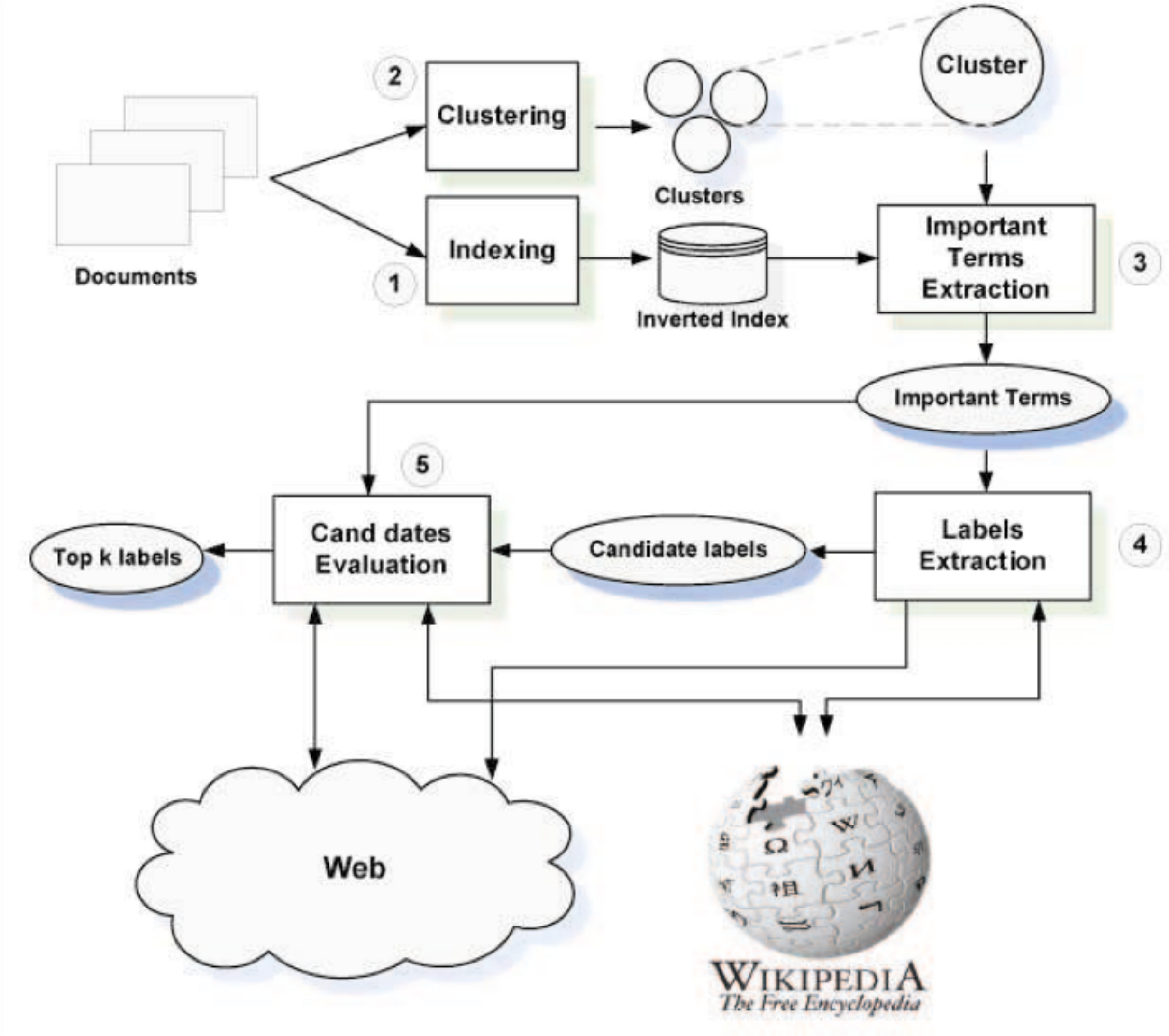
Why cluster documents

- hard to understand a large list of documents
- documents may be very different from each other

What do clusters mean?

- clustering algorithms group together similar documents
- how do we know which cluster to look in for the documents we want?
- need to meaningfully label clusters

General Framework



Indexing

- use R text mining package
  - R is a programming language and software environment for statistical computing and graphics

- use term-frequency in clustering and JSD

Clustering

- cluster using the Latent Dirichlet algorithm in R

- for each cluster
  - export the top N most representative words

- rest of the system implemented in python

Important Term Extraction

- reorder top terms
  - rank more informative terms higher
  - use Jenson Shannon Divergence

$$D_{JS}(P||Q) = \frac{1}{2}(D_{KL}(P||M) + D_{KL}(Q||M))$$

Label Extraction

- find Wikipedia documents relevant to cluster terms
    - search Yahoo for term disjunction
    - for each Wikipedia page returned
        - extract title and categories

Label Evaluation

- rank extracted labels by relevance
- multiple ways to judge relevance
  - mutual information
  - score propagation

Judging with Score Propagation

- score label using the rank of the documents it appears in
  - a label is better if it is in highly ranked documents

Review of Labeling Pipeline

- cluster and index documents
- extract important terms
  - re-score important terms
- extract labels from Wikipedia
  - using Yahoo! search
- score labels using search rank
- return top N labels

Experiments

- Corpus to cluster and label
  - 20 Newsgroups

- External corpus
  - Wikipedia
  - Yahoo! Search Results

Dataset - 20 Newsgroups

- 20,000 documents
- Average 1000 per news group

comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey sci.crypt
sci.electronics
sci.med
sci.space misc.forsale talk.politics.misc
talk.politics.guns
talk.politics.mideast talk.religion.misc
alt.atheism
soc.religion.christian

Evaluation

- Match@K
  - the number clusters with >= 1 correct top-K label

- MRR@K
  - the marginal reciprocal rank

- for K = 1 through 5

- same metric as used in Carmel et al.

Evaluation Details

- what's the ground truth?
  - any unabbreviated news group names
  - all stemmings
  - all synonyms (from WordNet)

- example

`sci.crypt` becomes `discipline, steganography, scientific, cryptography, science, cryptanalytics, coding, cryptology, cryptanalysis, secret, cryptographi, skill, writing, scienc`

Results - match@k

match@k	1	2	3	4	5
20 results, 20 terms	0.55	0.7	0.75	0.85	0.95
100 results, 20 terms	0.6	0.9	0.9	0.9	0.9
100 results, 20 terms - normalized	0.25	0.45	0.55	0.75	0.75
100 results, 20 terms, anchor tags	0.25	0.4	0.55	0.6	0.65

Results - mrr@k

mrr@k	1	2	3	4	5
20 results, 20 terms	0.55	0.625	0.642	0.67	0.69
100 results, 20 terms	0.6	0.75	0.75	0.75	0.75
100 results, 20 terms - normalized	0.25	0.35	0.38	0.43	0.43
100 results, 20 terms, anchor tags	0.25	0.325	0.375	0.3875	0.3975

Results

- using 20 top terms
- using 100 most relevant Wikipedia pages
- using score propagation

news groups	top 2 labels
comp.os.ms-windows.misc	American inventions, Tractor manufacturers of the United States
soc.religion.christian	People, Christian
sci.space	Science, Earth
rec.sport.baseball	Baseball, Ball
comp.sys.mac.hardware	Computer companies of the United States, Computer
?	States of the United States, Presidents of the United States
alt.atheism	Bible, Prophets of Islam
soc.religion.christian	Megafauna of Eurasia, Megafauna of Australia
sci.electronics	Email, American inventions
talk.politics.mideast	Lists of countries, European countries
rec.sport.hockey	Populated places established in 1793, Cities in Michigan
sci.crypt	Cryptography, Encryption
sci.med	Medicine, Health
talk.politics.mideast	Eastern Mediterranean countries, Middle Eastern countries
sci.crypt	Privacy, United States
comp.windows.x	American drama television series, 2000s American television series
talk.politics.misc	Presidents of the United States, President of the United States
rec.motorcycles	Motorcycle manufacturers of Japan, Tractor manufacturers of Japan
comp.graphics	Computer file, Computer file formats
comp.sys.ibm.pc.hardware	SCSI, Computer

Some Problems

- 1st cluster covers PGP encrypted documents, like:

```
------------ Part 7 of 14 ------------
MI:6EI:6EI:6EI:6EI:6EI:6EI:6EI4='1T='1T='1T='I:6EI:6EI:6EI0L+
M"PL+"PL+"PL+"PL+"PL+I:6EI:6E1T='1T='W]_?W]]Z>GHJ*BJ[NUY>`P,#
M`YF9F9D]/3T]/9F9F0->7KN[N[N[7EX#`P.9F9F'AX>'AX>'AX>'AX>'AX>'
...
```

- 6th cluster covers 3 news groups
  - comp.windows.x
  - rec.sport.hockey
  - talk.politics.guns

- commonalities based on headers and footers of messages

Parameters

- how many cluster terms to search with?
  - 20 avoids drift
- how many search result documents?
  - 100 gives good results, more is better
  - rank weighting controls quality
- how to do score propagation?
  - normalize key words?

Results

- using 20 top terms
- using 100 most relevant Wikipedia pages
- using score propagation with normalized keywords

news groups	top 2 labels
comp.os.ms-windows.misc	Acronyms, Quantum optics
soc.religion.christian	Judeo-Christian topics, Bible
sci.space	Toronto, Terrestrial planets
rec.sport.baseball	Olympic sports, Team
comp.sys.mac.hardware	Computers, Optical spectrum, Color, Shades of green, Apple Inc.
?	Information, States and territories established in 1816
alt.atheism	Bible, Judeo-Christian topics
soc.religion.christian	Religion, Atheism
sci.electronics	Email, Electronic documents
talk.politics.mideast	Republics, Jews
rec.sport.hockey	Underground Railroad locations
sci.crypt	Cryptography, Encryption
sci.med	Medicine, Oncology
talk.politics.mideast	Liberal democracies, Zionism
sci.crypt	Privacy, Identity management
comp.windows.x	Literary criticism, Author
talk.politics.misc	Titles, Management occupations
rec.motorcycles	Insurance, Institutional investors
comp.graphics	Lossy compression algorithms, JPEG
comp.sys.ibm.pc.hardware	SCSI, Macintosh internals

WikiCluster: Medicine, Health
Normalized: Medicine, Oncology

Original: Cramer, Medical

1st most representative doc

...

One of the responsibilities of a licensed physician is to read the medical
literature to keep up with changes in medical practice. All the clamor
over laetril resulted in the NCI spending quite a bit of money on clinical
trials

....

2nd most representative doc

...

A new natonal study on male sexual behavior, the most thorough examination of American men's sexual practices published since the Kinsey report more than four decades ago, shows about 2 percent percent of the men surveyed had engaged in homosexual sex

...

Clusters Overlap Newsgroups

- 2nd most relevant doc from talk.politics.misc
  - most frequent word, 'cramer', from here
  - email and name of user posting research studies on sexuality

- WikiCluster still finds more useful labels

Clusters Overlap Newsgroups

- Similar Newsgroups not divided as originally
  - Religion
    - talk.religion.misc
    - alt.atheism
    - soc.religion.christian
  - Windows
    - comp.os.ms-windows.misc
    - comp.windows.x
  - Politics
    - talk.politics.misc
    - talk.politics.guns

- This is a feature

WikiCluster: Cryptography, Encryption
Original: Clipper, Chip

1st most relevant document

...

One very interesting thing I notice about this is that the only use of the chip key is to encode the session key for the law enforcement field.

It is not used at all by the encryption algorithm.

...

2nd most relevant document

...

Once the session key K is established, the Clipper Chip is used to encrypt the conversation or message stream M (digitized voice). The telephone security device feeds K and M into the chip to produce two values

...

WikiCluster: Privacy, United States
Normed: Privacy, Identity management

Original: Government, Health

1st most relevant document

...

The initiative will involve the creation of new products to accelerate the development and use of advanced and secure telecommunications networks and wireless communications links

...

2nd most relevant document

...

It looks like Dorothy Denning's wrong-headed ideas have gotten to the Administration even sooner than we feared. It's time to make sure they hear the other side of the story, and hear it loudly!

...

Cluster Divide News Groups

- the previous cluster is about the *Clipper Chip* and related privacy issues
  - sub-division of sci.crypt news group
- original top words are misleading
- both WikiCluster and normalized WikiCluster produce understandable labels

Evaluation Problems

- our clusters aren't the newsgroup's clusters

- a high scoring label doesn't mean an intuitive label
  - and vice-versa

- clustering labeling solves a human interaction problem
  - human evaluation would be ideal

Extensions

- other sources of labels
  - anchor tags
  - related pages

- use web graph
  - weight with inlinks and outlinks

- reduce label redundancies

Labeling With Anchor Tags

- problem with duplicate terms in labels
- interesting that results are different and still understandable

news groups	top 2 labels
comp.os.ms-windows.misc	laser, ford laser
soc.religion.christian	jesus, bible
sci.space	science, earth science
rec.sport.baseball	baseball, league
comp.sys.mac.hardware	apple, apple computer
?	indiana indiana, indiana
alt.atheism	religion, bible
soc.religion.christian	religion, religion science
sci.electronics	email mail, microsoft mail
talk.politics.mideast	turkey, history turkey
rec.sport.hockey	toronto, city toronto
sci.crypt	security, security information
sci.med	cancer, history cancer
talk.politics.mideast	israel, history israel
sci.crypt	insurance, health insurance
comp.windows.x	files, email mail
talk.politics.misc	president, president united states
rec.motorcycles	insurance, history insurance
comp.graphics	color, jpeg compression
comp.sys.ibm.pc.hardware	computer, computer software

Questions?

Resources and References

- 20 Newsgroups
- WordNet
- Yahoo! Search Boss
- David Carmel, Haggai Roitman, Naama Zwerdling: Enhancing cluster labeling using Wikipedia.SIGIR 2009: 139-146
- David Carmel, Elad Yom-Tov, Adam Darlow, Dan Pelleg: What makes a query difficult?. SIGIR 2006: 390-397