Discovering Generalized Concepts with a Category Graph

Tom Vacek
with John Joseph and Dan Boley
University of Minnesota

Text Mining Workshop 2010

Supported by:
NSF grant 0534286 & DARPA STTR W31P4Q-08-C-0242
Goal

- Discover document topics at varying levels of generality.
- Use domain knowledge to interpret.
- Find concise labels for these topics.
What is a Topic?

A concise phrase describing what the document discusses.

Topic should be concise label rather than a collection of related words.
Outline

Finding concepts with Wikipedia overview
Our approach
Results/conclusions
Bowman tended to overplay Francis at times because he is a Bowman-style player. He plays hard at all times, doesn't disregard his defensive responsibilities and is a good leader. Bowman rewarded him by increasing his ice time.

Jagr can be very arrogant and juvenile and display a "me first" attitude. This rubbed Bowman the wrong way and caused him to lose some ice time.

Throughout the year, Francis consistently received more ice time than Jagr. Although I have never seen stats on this subject, I am pretty sure that Jagr had more points per minute played than Francis. When you add to that Jagr's better +/- rating, I think it becomes evident that Jagr had a better season - not that Francis had a bad one.

Gregmeister
Oil prices spike as hurricane threatens Gulf.
ND ethanol plant files for bankruptcy.
Wikipedia for finding topics

- Wikipedia:
  - $10^6$ articles
  - $10^5$ categories
- Articles tagged with categories
- Category structure forms directed graph
- Idea: Do term assignment using Wikipedia categories
Problems

- **Scalability**: $10^5$ classes, $10^6$ training docs
  - Feature selection: $2^{35}$ scores required

- **Precision problem**: too many output classes
  - A random document is in a given class with extremely small probability.
  - ”Needle in haystack”
Precision Problem (1/3)

Assume: 1) Classifier g is a binary classifier that recognizes some class c.

2) Classifier g has 99.9% recall accuracy on c and its complement.

3) Document occurs in c with $10^{-5}$ probability.

What is precision of g?
Precision Problem (2/3)

Precision = P(C|+)

'C' = 'd is in c'

'+' = 'g has positive prediction'

Bayes' Theorem:

\[
P(C|+) = \frac{P(+|C) \cdot P(C)}{P(+)} = \frac{\cdot999 \cdot 10^{-5}}{.999 \cdot 10^{-5} + (1 - .999)(1 - 10^{-5})} = 0.009891
\]

< 1\%
Precision-recall brittleness for a class with $10^{-5}$ prior probability
Related Work

- Coursey, Mihalcea, Moen (2009)
  - Max precision: .16 @ 5 categories on test subset of Wikipedia.

- Schonhofen (2006)
  - Trains on W. article titles only.
  - Precision: .52 @ 1, .31 @ 5 categories.
  - Train/test set overlap?
Outline

Finding concepts with Wikipedia overview

Our approach

Results/conclusions
Our Approach

Use a naive but efficient classifier.

Use the category graph structure of Wikipedia to de-noise the results.

Relax expectations.

- Find concepts that have general (man-on-the-street) context.
Rocchio Classification
Concept Space Transform

- Document transformed into vector of affinities with each Wikipedia category.
- Called 'base scores'
De-noise results
Random Walk

- $\beta$ is a user parameter for bias the random walk
- At node $i$, take a step
  - Toward a parent category $u$ of $i$ with $\beta$ probability.
  - Toward a child category $v$ of $i$ with probability $(1-\beta)$
  - Teleport to a random node with small probability
- Intuition: find nodes that have lots of edges to high-scoring neighbors.
Random Walk

- $s_j$ is base score of node $j$
- $F$: parent-step matrix. $F_{ij} = s_j$ if $j$ is parent of $i$
- $R$: child-step $R_{ij} = s_j$ if $i$ is parent of $j$.
- $S$ is normalized vector of base scores.
- Empty rows get $s$. $F$, $R$ row normalized.
- Stationary distribution is leading left e-vector:
  $\alpha(\beta F + (1-\beta)R) + (1-\alpha)(1...1)^T s^T$
- Called 'generalized topic'
Outline

Finding concepts with Wikipedia overview
Our approach
Results/conclusions
Subject: stats for hockey

I'm the keeper of the stats for a family hockey pool and I'm looking for daily/weekly email servers for playoff stats. I've connected with the servers at J.Militzok@skidmore.EDU and wilson@cs.ucf.edu. I'm still sorting these two out. Are there others? Email please as my site doesn't get this group. Thanks. Rick -- rick@emma.panam.wimsey.bc.ca
It's great that all these other cars can out-handle, out-corner, and out-accelerate an Integra. But, you've got to ask yourself one question: do all these other cars have a moonroof with a sliding sunshade? No wimpy pop-up sunroofs or power sliding roofs that are opaque. A moonroof that can be opened to the air, closed to let just light in, or shaded so that nothing comes in.

You've just got to know what's important.

--Doug '93 Integra GS
Subject: last

I have two questions: 1) I have been having troubles with my Wordperfect for Windows. When I try to select and change fonts, etc. some of the text disappears. I tried to center two lines once, and the second line disappeared. I can not find the error, and I do not know how to correct it. 2) Is this the right newsgroup? Where should I go?

E-mail prefered...

Who else is still waiting for "Naked Gun Part (Pi)"?
I sometimes see OTC preparations for muscle aches/back aches that combine aspirin with a diuretic. The idea seems to be to reduce inflammation by getting rid of fluid. Does this actually work?

Thanks,

--Larry C.
Subject: Re: Essene New Testament

[William Christie asked about the Essene NT. Andrew Kille responded: There is a collection of gospels which usually goes under the name of the "Essene Gospel of Peace." These are derived from the gnostics, not the essenes, and are ostensibly translations from syriac texts of the fourth and fifth centuries (I vaguely recall; I can't find my copy right now).

--clh]

There had been recent criticism of this in a listserv for academic Biblical scholars: they all say the book(s) are modern fakes.

D.H.
<table>
<thead>
<tr>
<th>Topic</th>
<th>Base</th>
<th>Gen</th>
<th>Base</th>
<th>Gen</th>
<th>Base</th>
<th>Gen</th>
<th>Base</th>
<th>Gen</th>
<th>Base</th>
<th>Gen</th>
<th>Base</th>
<th>Gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>rec.sports.baseball</td>
<td>baseball pitching</td>
<td>79</td>
<td>computer keys</td>
<td>128</td>
<td>cryptography</td>
<td>273</td>
<td>earth orbits</td>
<td>66</td>
<td>heavy machine guns</td>
<td>81</td>
<td>cities and towns in armenia</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>baseball computer games</td>
<td>71</td>
<td>sound chips</td>
<td>82</td>
<td>clippers</td>
<td>70</td>
<td>space shuttle program</td>
<td>58</td>
<td>wildfires</td>
<td>54</td>
<td>jews by country</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>chicago cubs</td>
<td>26</td>
<td>block ciphers</td>
<td>57</td>
<td>algorithms</td>
<td>58</td>
<td>space advocacy</td>
<td>47</td>
<td>firearm laws</td>
<td>40</td>
<td>political parties in palestine</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>major league designated hitters</td>
<td>25</td>
<td>law</td>
<td>43</td>
<td>law</td>
<td>34</td>
<td>gamma-ray telescopes</td>
<td>25</td>
<td>bill clinton</td>
<td>24</td>
<td>israeli-palestinian conflict</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>soc.religion.christian</td>
<td>singular god</td>
<td>68</td>
<td>cryptography</td>
<td>273</td>
<td>earth orbits</td>
<td>66</td>
<td>space</td>
<td>150</td>
<td>firearms</td>
<td>181</td>
<td>judaic studies in academia</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>christian viewpoints</td>
<td>65</td>
<td>clippers</td>
<td>70</td>
<td>space exploration</td>
<td>143</td>
<td>space exploration</td>
<td>143</td>
<td>weapons</td>
<td>64</td>
<td>history of armenia</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>jesus</td>
<td>56</td>
<td>algorithms</td>
<td>58</td>
<td>astronomy</td>
<td>69</td>
<td>hip objects</td>
<td>16</td>
<td>firefighting</td>
<td>49</td>
<td>cyprus</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>study bibles</td>
<td>45</td>
<td>law</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>christianity</td>
<td>307</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>religious faiths, traditions, and movements</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>jesus</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lgbt</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sci.sports.hockey</td>
<td>defunct ice hockey leagues</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>boston bruins coaches</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>power rangers</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>toronto maple leafs coaches</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ice hockey</td>
<td>191</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>computer and video games</td>
<td>179</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sports</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nhl</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rec.autos</td>
<td>new york city subway passenger equipment</td>
<td>234</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>oils</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>insurance companies of japan</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>automotive braking technologies</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>automobiles</td>
<td>433</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>engines</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>oils</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>vehicles</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
News Group Results

![Graph showing the number of concepts versus percent for different groups](image)
Conclusion

- Generalized topics are more homogeneous.
- Generalized topics have useful context; they reveal news group names.
- Generalized topics seem more germane than base assignments, but also more general.
Related Work

- Coursey, Mihalcea, Moen (2009)
Related Work

- Schonhofen (2006)
- Trains on titles only
- Overlap in test/training sets

![Graph showing precision and recall of category selection](image)