Computer-Assisted Keyword and Document Set Discovery from Unstructured Text

Gary King[1] Patrick Lam[1] Margaret E. Roberts[2]

[1]Institute for Quantitative Social Science
Harvard University

[2]University of California, San Diego

Slides prepared for PolMeth XXXI at the University of Georgia
July 24, 2014

An Essential Component of Text Analysis: Keywords

1. Define a corpus of documents. ? (often via Boolean keyword search)
2. Apply sophisticated text analysis methods. ✓
3. Get substantive results. ✓

Example: Studying tweets related to the Boston Marathon bombings

1. Start with all tweets containing *boston*.
2. Discard tweets containing *red & sox*, *bruins*, *celtics*.
3. Search for tweets containing the words *#bostonmarathon*, *suspect*, *tsarnaev*, *dzhokhar*, *explosion*, *terrorism*.

Using a reasonable set of keywords may miss a lot of relevant documents.

What We May Have Missed

Boston Bombings

EVENT	SUSPECTS	VICTIMS	REACTION	POLITICAL
explosion	tsarnaev	innocent	tragedy	obama
terrorism	dzhokhar	victim	prayers	#tcot
attack	tamerlan	collier	#prayforboston	#benghazi
⋮	⋮	⋮	⋮	⋮

How do we find enough keywords that will capture the concept of interest?

Keywords in Political Science

Boolean keyword search is used for selecting topics in...

- newspaper articles (Ho and Quinn 2008; Eshbaugh-Soha 2010; Gentzkow and Shapiro 2010; Puglisi and Snyder 2011)
- social media (Hopkins and King 2010; King, Pan and Roberts 2013; Jamal et. al. 2014)
- court cases (Gill and Hall 2013)
- congressional bills (Kim 2014)

...but keywords are also useful in other ways.

Why We Need Keywords: Conversations Evolve

- Social trends
  - Twitter hashtags: "#BostonBombings" $\rightsquigarrow$ "#PrayforBoston"
- Political positioning
  - "late term abortion" $\rightsquigarrow$ "partial birth abortion"
  - "pro-choice v pro-life" $\rightsquigarrow$ "reproductive rights"
- Evading (law enforcement) detection
  - child pornographers using different terms to evade detection

How do we find keywords to follow these conversations?

Evading Censorship in Chinese Social Media

Example Substitution 1: Homograph

自由 "Freedom" 

目田 "Eye field" (nonsensical)

Example Substitution 2: Homophone (both sound like "hexie")

和谐 "Harmonious [Society]" (official slogan) 

河蟹 "River crab" (irrelevant)

How do we find the keywords to follow this conversation?

The current technology for finding keywords. . .

How good are humans at thinking of keywords?

A Small Formal Experiment

We asked 43 undergrads to think of keywords about Obamacare and Boston marathon bombings:

> *We have 10,000 Twitter posts, each containing the word "healthcare" from the time period surrounding the Supreme Court decision on Obamacare. Please list any keywords which come to mind that will select posts in this set related to Obamacare and will not select posts unrelated to Obamacare.*

Results: Humans are Unreliable and Limited

Obamacare

Boston bombings

- Unreliable: 66% and 59% of the words were suggested by only 1 out of 43 people.
- Limited: Median number of words per respondent was 8 and 7.

Human Scorecard

GOOD	BAD
• recalling a small list of good keywords • recognizing many good keywords when they see it	• recalling the same list of keywords every time (unreliability) • recalling a long list of keywords that capture different ways of representing a concept: "part-list cuing"

We need a keyword discovery method that takes advantage of what humans do best and helps humans with what they do poorly.

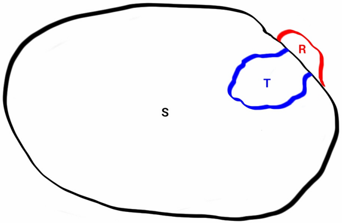
Existing Options for Automated Keyword Discovery

- Search queries or your web log (Google Adwords)
  - requires structured data
- Thesaurus methods (reference books, wordnet, etc)
  - requires a relevant thesaurus (that follows current and future trends in language)
- Co-occurrence methods
  - requires the documents to contain the original keyword

We need a new computer-assisted keyword discovery method that...

- requires only unstructured text (but can use other information if available)
- mines even from documents not containing original keyword
- works with novel words in *any* language
- helps humans find more and better keywords faster

Setting Up the Algorithm



- Reference set R: documents about a concept of interest (selected by methods that humans are good at)
  1. hand-select documents
  2. select small number of high quality keywords to search for documents
  3. good existing set
- Search set S: broad set of documents
- Target set T: documents in S about same concept as in R

- The Goal: Find keywords K_T (Boolean operators) that define Target set
- Keep humans in the loop: users decide which keywords to choose

Keyword Discovery Algorithm (simplified)

Part 1: Using Classifier "Mistakes" to Find Target Set

1. Create training set of docs by randomly drawing from R and S
2. Fit classifier to training set
3. Classify each S doc into R or S using classifier fit

		Classification	
		Search	Reference
Truth	Search	$\{S\|S\}$	$\{R\|S\}$
	Reference	$\{S\|R\}$	$\{R\|R\}$

(mis)classified search doc into reference set

4. Define as Target Set the documents which are classified as $\{R|S\}$: classifier "mistakes"

Keyword Discovery Algorithm (simplified)

Part 2: Extracting Keywords from Target Set

- Define a rule r as a keyword or full Boolean term
  - single keyword
  - & statements: Gary & King & Harvard
- Collect list of candidate rules (the parameter space)
  - ideal but infeasible: all possible rules
  - practical: all common rules (via "Apriori" algorithm)
- Rank rules by how well they characterize the target set
  - rule is "good" if it appears in more docs in the target set than outside it
  - ceteris paribus, rules appearing more often in the target set should be ranked higher
  - metric: likelihood value from Beta-Binomial likelihood

$$L(r|\alpha, y_1, \ldots, y_n) = BB(n_{r,t}|n_r, \alpha) \times BB(n_{-r,t}|n_{-r}, \alpha)$$

- Users choose which keywords they like

Keyword Discovery Algorithm (extended version)

- Use multiple classifiers
  - different classifiers may have different "opinions" about documents because they capture different aspects of the documents
  - use classifier probabilities instead of discrete classifications
- Cluster documents by vector of classification probabilities
  - a way to combine the diverse classifier opinions
  - one or more clusters may approximate the target set
  - use as a keyword presentation method by grouping similarly classified documents together to help provide context
  - may possibly pick up different concepts in search set
- Choose rules that characterize each cluster well
  - rank rules by how well a rule classifies a cluster versus rest of the search set

Validating the Algorithm

- Validation for keywords:
  - user decides if keyword is helpful or interesting
  - user can search and read documents containing the keywords
- Validation for target set retrieval with keywords
  - difficult without knowing target set
  - approximate target set with Twitter hashtags (self-coded topics)
  - example: Mandela's passing (topic hashtag: #Madiba)
  - metrics:
    - recall: $\frac{\#\text{ of target set documents retrieved by keyword(s)}}{\#\text{ of documents in target set}} \times 100$
    - precision: $\frac{\#\text{ of target set documents retrieved by keyword(s)}}{\#\text{ of documents retrieved by keyword(s)}} \times 100$

Following Nelson Mandela's passing

R: Mandela & #Madiba
T: #Madiba & NOT Mandela
S: *T* + South Africa & NOT Mandela

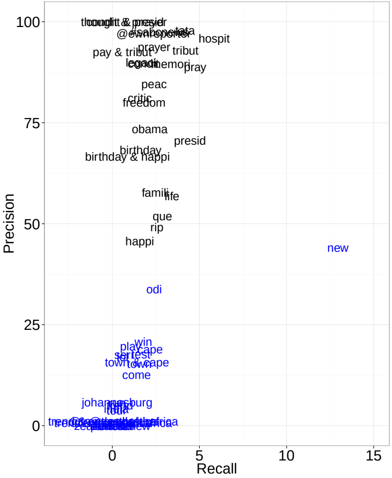
Cluster 1 (Mandela)
(closer to reference set)

tribute, prayer, president, family, birthday, hospitalized, tata, freedom, pray, happy & birthday, happy, life, critical, condition, peace, @ewnreport, rip, memory, #sabcnews, thoughts & prayers, obama, legacy, pay & tribute, condition & president

Cluster 2 (South Africa)
(further away from reference set)

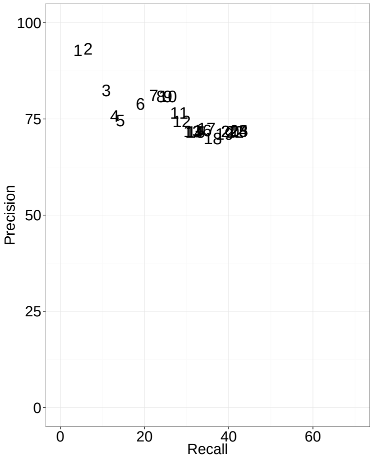
trend, cape, town, cape & town, come, @trendssthafrica, trend & @trendssthafrica, pakistan, new, test, india, tour, australia, run, win, zealand, new & zealand, johannesburg, odi, lol, tressthafrica, trend & trendssthafrica, series, #cricket, play

Algorithm Separates Relevant and Irrelevant Words



- Mandela keywords have high precision
- South Africa keywords have lower precision
- Low recall in general due to diversity of keywords used
- Increase recall by taking the union of multiple Mandela keywords (OR rules): *tribute* OR *prayer*

Increase Recall with OR Rules



- Each number represents the number of keywords combined with “OR” starting from the top of Mandela cluster
  - ex: “3” represents the rule *tribute* OR *prayer* OR *president*
- Recall monotonically increases with each added keyword
- Precision may increase or decrease
- Extend algorithm by ranking OR rules as well

Following the Obamacare Supreme Court Case

R: Obamacare

S: healthcare & NOT Obamacare

Cluster 1 (Obamacare)
(closer to reference set)

supreme, court, constitutional, obama, mandate, law, uphold, president, republican, congress, roberts, senate, repeal, insurance, tax, rule, justice, affordable, decision, penalty, unconstitutional, hill, act, clause, commerce

Cluster 2 (general healthcare)
(further away from reference set)

inform, manage, develop, medicine, intern, help, learn, train experiment, study, resource, city, industries, operate, hospital, build, time, food, clinic, receive, medical, profession, research, during, data

Following the Boston Bombings in Social Media

R: #BostonBombings
S: Boston & NOT #BostonBombings

Cluster 1 (bombings)
(closer to reference set)

suspect, police, people, fbi, suspect & marathon, report, terror, tsarnaev & suspect, police & suspect, investigate, news, arrest, tsarnaev, kill, muslim, obama, cnn, #bostonmarathon, dzhokhar, terrorist, #prayforboston, #tcot, #benghazi

Cluster 2 (sports)
(further away from reference set)

game, red, sox, red & sox, celtics, bruins, fan, tonight, back, come, win, play, chicago, love, new & york, #mlb, team, series

Finding Conversation about Your Retirement Savings

R: "save for retirement"

S: retirement

Cluster 1 (Your retirement savings)
(closer to reference set)

savings, income, financial, money, invest, tax, debt, fund, pay, amount, payment, account, rate, expensive, ira, financing, asset, plan, cost, loan

Cluster 2 (Sports star retirement)
(further away from reference set)

team, sport, announcement, star, player, former, fan, league, game, season, man, play, football, win, champion, club, championship, city, saturday, night

The Bo Xilai Scandal in China

R: Bo Xilai 薄熙 来

S: Chongqing 重庆 (City where Bo was mayor) & NOT Bo Xilai

王立军	Wang Lijun (Chongqing police officer, fled to U.S. consulate)
政治	government
事件	[Chongqing] event (euphemism for "Bo Xilai scandal")
打黑	strike corruption
犯罪	commit a crime
民主	democracy
权力	power
文革	Cultural Revolution
领导	leader
改革	reform
群众	the masses
中央中共	Central Communist Party
社会主义	socialism
唱红	sing red songs
黑社会	black society
干部	cadre
路线	party line

Finding Writings about Suicide Bombings

R: "martyrdom operations" عمليات الاستشهادية from "Haqibat al-Mujahid"

S: the Jihadist library ("Pulpit of Tawhid and Jihad")

العدو	enemy
قتل	killing
والنكاية	to vex or spite ("vex the infidels")
يَعْلَمُهُمْ	teach them
الْخَيْلِ	steed
وَأَعِدُّوا	fight
تُظْلَمُونَ	wronged
ترهبون	terrify
الغلام	boy (refers to the story of the boy and the king, relevant to jihadis)

Quran 8:60

> *And prepare against them whatever you are able of power and of steeds of war by which you may terrify the enemy of Allah and your enemy and others besides them whom you do not know [but] whom Allah knows. And whatever you spend in the cause of Allah will be fully repaid to you, and you will not be wronged.*

For more information

http://j.mp/wordstakes

GaryKing.org
PatrickLam.org
j.mp/MollyRoberts

Beta-Binomial Likelihood

Define:

- cluster of interest as c and rest of S as $-c$
- $y_d = 1$ if doc $d \in c$ and $y_d = 0$ if $d \in -c$
- $n_{r,c}$ as the $\#$ of docs in c that contain r
- $n_{-r,c}$ as the $\#$ of docs in c that do not contain r
- n_r and n_{-r} as $\#$ of docs in S that contain and do not contain r

Likelihood:

$$L(r|\alpha, y_1, \ldots, y_n) = BB(n_{r,c}|n_r, \alpha) \times BB(n_{-r,c}|n_{-r}, \alpha)$$

Non-identification due to symmetry:

- rules that have high likelihood can characterize either c or $-c$
- look at percentage of documents in c and $-c$ that contain r
- if higher percentage in $-c$, drop r or change r to a NOT rule for c