Introduction
PageRank and fluency
Dual-route topic model

Cognitive Modeling

Lecture 13: Memory and Information Retrieval

Sharon Goldwater

School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk

February 25, 2010

Introduction
PageRank and fluency
Dual-route topic model

Reading: Steyvers and Griffitths (2008).

Introduction
PageRank and fluency
Dual-route topic model

Motivation
Rational models of memory

Motivation

Why more about memory? Why Steyvers and Griffitths (2008)?

- Example of very recent work, useful to see what's happening now and compare to older work.
- Shows that there can be synergies between cognitive modeling and engineering applications.
- Includes quantitative (not just qualitative) comparisons to other models.

Introduction
PageRank and fluency
Dual-route topic model

Motivation
Rational models of memory

Rational models of memory

Anderson (1990) proposed that the need-probability p of an item A depends on its *history of use* H_A and the set of *contextual cues* Q that are present:

$$p \quad \propto \quad P(A|H_A)P(Q|A)$$

- Model was based on models of library usage.
- Can newer information retrieval systems provide additional insight?

Introduction
PageRank and fluency
Dual-route topic model

Motivation
Rational models of memory

Steyvers and Griffiths (2008)

Steyvers and Griffitths (2008) make the analogy between human memory and computer information retrieval.

$$P(d|q) \propto P(q|d)P(d)$$

- document (item(s) in memory)
- query (memory probe)

Paper considers $P(d)$ first, then $P(q|d)$.

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Information retrieval

Simple information retrieval system:

1. Collect all pages containing the query words q.
2. Rank pages by importance.

Viewed as Bayesian model:

$$P(q|d) = \begin{cases} 1 & \text{if } q \in d \\ 0 & \text{otherwise} \end{cases}$$

so that $P(d|q) \propto P(d)$.

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

PageRank

How to determine the importance of a document $P(d)$?

Google's PageRank method (Brin and Page, 1998) assumes a recursive definition:

- Important pages have links from many other pages.
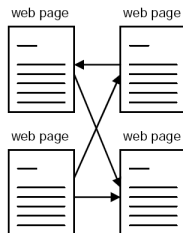- Important pages have links from other important pages.

Figure from Steyvers and Griffitths (2008)

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

PageRank

PageRank viewed as *random surfing*:

- Surfer begins at a random web page
- Follows a link from that page uniformly at random
- Continues this procedure forever (assuming web doesn't change meanwhile).

A document's PageRank is proportional to the number of times the surfer visits it. (Based on theory of Markov chains)

PageRank can also be easily computed in fixed time using linear algebra.

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Semantic networks

Many psychologists believe words/concepts are represented in the mind as a *network of associations*:
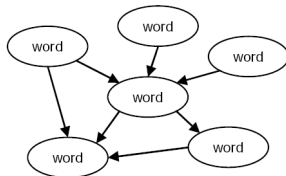
- Similar to network of web pages.
- Might "important" concepts be defined similarly to important web pages?

Figure from Steyvers and Griffitths (2008)

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Fluency Task

To test this idea, we need a task that focuses only on the prior probability of an item, i.e. where $P(q|d)$ is either 0 or 1 (as in info retrieval).

- Steyvers and Griffiths (2008) asked subjects to "think of a word beginning with the letter..."
- q is the letter prompt, d is the word response.
- This is known as a *fluency* test, sometimes used for diagnosing disorders.

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Fluency Data

Results for various letters, pooled across 50 subjects:

		Beginning letter		
A	B	C	D	P
		Human responses		
Apple (25)	Boy (11)	Cat (26)	Dog (19)	People (5)
Alphabet (7)	Bat (6)	Car (8)	Dad (16)	Penguin (3)
Ant (6)	Banana (5)	Cool (3)	Door (5)	Pizza (3)
Aardvark (3)	Balloon (4)	Card (2)	Down (4)	Play (3)
Ace (2)	Book (4)	Class (2)	Dark (3)	Pop (3)
Ambulance (2)	Baby (3)	Coke (2)	Dumb (3)	Puppy (3)
Animal (2)	Ball (2)	Cookie (2)	Day (2)	Piano (2)
Absence (1)	Barn (2)	Crack (2)	Devil (2)	Pie (2)
Acrobat (1)	Bear (2)	Cross (2)	Dinosaur (2)	Pig (2)
Act (1)	Beef (2)	Cut (2)	Do (2)	Power (2)

Figure from Griffiths et al. (2007)

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Link structure

Testing hypothesis also requires knowing the link structure of the semantic network. Steyvers and Griffitths (2008) used existing *word association data*.

- Task: given a word prompt (e.g. *dog*, *blue*), name the first word that comes to mind (*cat*, *yellow*).
- Responses collected for many subjects, association data includes 5k words.

Steyvers and Griffitths (2008) assume a semantic link from each word to its associates.

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Model evaluation

Steyvers and Griffitths (2008) consider two alternative ways to compute a word's importance:

- **word frequency**: computed from a very large corpus. Similar to Anderson's (1990) history model but without spacing.
- **in-degree**: the number of links pointing to this word. Similar to PageRank but with all links weighted equally.

Each method is used to rank words by importance.

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Rankings

Top-ranked words by humans and models for letter *d*:

Human responses		PageRank		In-degree		KF Frequency	
DOG	(19)	DOG	(19)	DOG	(19)	DO	(2)
DAD	(16)	DARK	(3)	DEATH	(1)	DOWN	(4)
DOOR	(5)	DRINK	(1)	DRINK	(1)	DAY	(2)
DOWN	(4)	DOWN	(4)	DIRTY	(0)	DEVELOPMENT	(0)
DARK	(3)	DEATH	(1)	DARK	(3)	DONE	(1)
DUMB	(3)	DOOR	(5)	DOWN	(4)	DIFFERENT	(0)
DAY	(2)	DAY	(2)	DIRT	(0)	DOOR	(5)
DEVIL	(2)	DIRTY	(0)	DEAD	(0)	DEATH	(1)
DINOSAUR	(2)	DIRTY	(0)	DANCE	(0)	DEPARTMENT	(0)
DO	(2)	DEAD	(0)	DANGER	(1)	DARK	(3)

Figure from Steyvers and Griffitths (2008)

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Quantitative comparison

How far down the model's list (in %) before 50% of human
responses are covered?

Predictor	All words	Nouns only	Concrete nouns only
PageRank	8.33	8.16	13.33
In-degree	10.00	14.77	17.54
Word frequency	29.09	36.54	13.33

PageRank is a significantly better predictor than frequency or
in-degree.

Results from Griffiths et al. (2007)

Introduction
PageRank and fluency
Dual-route topic model

PageRank
Semantic networks
Human data
Results

Discussion

- Quantitative comparison shows PageRank predicts fluency better than previous proposals.

- Provides evidence both for semantic network representation and PageRank-like "importance" of concepts, showing connections between human memory and information retrieval.

- Does not explain why nouns are preferred responses.

- Does not explain why semantic network representation should exist.

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Memory test

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Verbatim and gist information

Psychologists have proposed two kinds of memory encoding.

- **Verbatim:** What animal is present? What color is the door? What is draped on the wheel? How many windows are there? What is on the frontmost windowsill?

- **Gist:** What kind of place is this? Does it seem old or modern? Quiet or busy? What are the buildings made of? Are there any people?

Similarly for other groups of items in memory (e.g., word lists).

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Rational analysis

Why have two kinds of encoding?

- Gist is useful: requires less memory (fewer resources), captures "aboutness", which helps for generalizing to new situations.
- Verbatim is useful: specific details could be important.

The same trade-off occurs in information retrieval:

- Search for *rational analysis cognitive*, or
- Search for *rational analysis Griffiths*

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Psychological data

When lists contain many related words, subjects falsely recall
"lure" words:

<div>

nail
tool
wrench   ⇒   hammer
wood
...

</div>

When lists contain many related words and one unrelated word,
subjects have better recall for the semantic isolate:

<div>

carrot
spinach
hammer   ⇒   hammer
eggplant
...

</div>

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Verbatim and gist models

To store a document (word list) verbatim, encode each item.

- Requires n memory chunks for n items.
- Model is basically a unigram word model.

To store the gist of a document, use a *topic model*.

- Each document is stored as a mixture of topics.
- May require only a few topics (i.e., a few memory chunks).

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Topic model

- Topic model assumes a fixed number of topics $t_1 \ldots t_N$.
- Each topic assigns probabilities to all vocabulary words:

t_1		t_2		t_3	
research	0.12	class	0.11	tango	0.12
model	0.08	student	0.083	music	0.08
grant	0.052	slides	0.041	dance	0.052
student	0.008	model	0.009	Argentine	0.008
segmentation	0.003	assignment	0.002	class	0.003
. . .		. . .		. . .	
tango	10^{-6}	tango	10^{-6}	grant	10^{-7}

- Each document is a mixture of topics.

$$d_1 = (.7, .25, .05), \ d_2 = (.01, .01, .98), \ d_3 = (.4, .5, .1), \ldots$$

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Putting it together

Dual-route model assumes each word is encoded either verbatim
(with prob. $P(v|d)$) or using gist model (with prob. $1 - P(v|d)$).

- Dual model infers probabilities of each route for each
  document, optimizing encoding.
- Model priors prefer using relatively few topics for encoding.
- Lists of related words are efficiently encoded using gist.
- Lists with few related words cannot be efficiently encoded
  using gist, so verbatim used instead.
- Semantic isolates also encoded more efficiently using verbatim.

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Results

Steyvers and Griffitths (2008) learn 1500 topics from a very large corpus of educational materials, then apply model to word lists from psychological experiments.

- When a list contains a semantic isolate, $P(\text{verbatim})$ is higher for that word in that list.
- As a result, probability of recall is higher for these words.
- When a list contains many related words, $P(\text{gist})$ is higher for that list, with a small number of topics having high probability.
- As a result, probability of recalling unseen words from the same topic(s) is higher.

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Discussion

- Rational analysis suggests usefulness of both verbatim and gist encoding in memory.
- Dual-route topic model shows how optimizing document probabilities automatically trades off between the two.
- DRTM exhibits semantic isolation and false memory effects.
- Implementing DRTM for information retrieval also improves performance.

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

Summary

Considering similarities between human memory and information retrieval can yield better cognitive models and better applications.

- PageRank works for Google, also models fluency better than previous models.
- Dual-route topic model explains semantic isolation and false memory effects, also improves information retrieval.

Outstanding issues:

- How do these two models relate to each other, and to Anderson's model?
- Can we develop a single unified model of memory?

Introduction
PageRank and fluency
Dual-route topic model

Verbatim and gist
Dual-route topic model
Evaluation

References

Anderson, John R. 1990. *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Griffiths, T., M. Steyvers, and A. Firl. 2007. Google and the mind: Predicting fluency with PageRank. *Psychological Science* 18:1069–1076.

Steyvers, Mark and Thomas L. Griffitths. 2008. Rational analysis as a link between human memory and information retrieval. In In M. Oaksford and N. Chater, editors, *The probabilistic mind: Prospects for rational models of cognition*, Oxford University Press, Oxford.