Ocelot: A system for summarizing web pages

Adam Berger

March 6, 2000

Carnegie Mellon
School of Computer Science

Why summarize web pages?

1. User aid in web search:
   1. UCI Machine Learning Group - **Machine learning** investigates the mechanisms by which knowledge is acquired through experience. Research at UCI spans the spectrum of models for **learning**, including those based on
      http://www.ics.uci.edu/AI/ML/Machine-Learning.html
   2. Online Machine Learning Resources - It is far from complete and is updated on an irregular basis. Please direct comments / suggestions / ... to Gerhard Widmer (gerhard@ai.univie.ac.at) Links 2 Go Key Resource **Machine**
      http://www.ai.univie.ac.at/oefai/ml/ml-resources.html
   3. David W. Aha: Machine Learning Page - **Machine Learning** Resources Suggestions welcome Applications | Bibliographies | Books | Commercial | Conference Announcements | Courses | Data Repositories | Home Pages |
      http://www.aic.nrl.navy.mil/~aha/research/machine-learning.html
   4. UCI Machine Learning Repository - This is a repository of databases, domain theories and data generators that are used by the **machine learning** community for the empirical analysis of **machine learning** algorithms.
      http://www.ics.uci.edu/~mlearn/MLRepository.html

• "If we can summarize web pages, we can summarize anything." (automatic title/abstract generation, "executive summary" of news headlines...)

Extractive summarization

Pick out phrases, sentences, even paragraphs:

blah blah blah blah blah blah
blah blah blah blah blah blah
blah blah blah blah blah blah
blah blah blah blah blah blah
blah blah blah blah blah blah
blah blah blah blah blah blah
blah blah blah blah blah blah
blah blah blah

- **Identify cue phrases**: "*Most importantly, …*"
- **Information extraction**: Distill out the "W"s.
- **Orthogonality**: Extracted parts shouldn't overlap in content.

Extractive summarization

Long, "well-behaved" text, like news articles, are often amenable to extraction.

TOKYO (AP) -- Sony's Playstation 2 hit stores in Japan on Saturday and sold out promptly, whisked off by eager fans who spent the night waiting in line to snap up the video-game machines.

Competition for the console, which sold for the quivalent of $370, was fierce in the Tokyo neighborhood of Akihabara, the nation's hub for electronics retailers.

Thousands of people were already lining up outside stores on Friday afternoon, hoping to buy the ...

Summarizing web pages often requires *synthesis*, not merely extraction....

Summarization, The YAHOO! Way

Hire a few thousand librarians

Mandarin Speaking Surfer
(Location - Santa Clara, CA)
Surf the Net for Yahoo! Chinese. Review Web sites for inclusion in the Chinese database, categorize them appropriately in the existing classification scheme, and help determine the evolution of that scheme. A passion for the Internet, careful attention to detail, an abundance of common sense, and excellent organizational skills are essential to meet the challenges of classifying an incredibly diverse array of unstructured information intuitively and efficiently. BA/BS or equivalent preferred. Fluency in both spoken and written Chinese (Mandarin) and English is a must. Cover letter required.
Interested? E-mail your resume and cover letter to jobs-chinese-surf@yahoo-inc.com *and note job code* **WEBSRFMAN**.

Summarizing web pages (without a $83B market cap.)

1. Start with a few thousand summarized web pages.
2. Let a computer "learn" (by inspection) how to automatically summarize web pages.

OpenDirectory Project

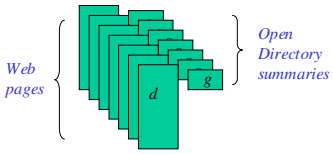
1,551,990 sites*, each summarized by a human editor

** and growing*

- ABWAG - Tiny Ron, an actor, created abwag based on teacher Don Richardson's feel, think, do technique in acting.
- 101% Acting - Directory of Acting sites with tips, how to's, news, history, actor bios, acting reference guides and more.
- Acting is Storytelling -For the New Age Actor - Acting, is, always has been and always will be, storytelling. The Millennium System teaches you how. Created by long time actor, Ken Farmer.
- Acting Workshop On-line (AWOL) - A resource for beginning actors, with practical info. Features articles and links.
- Acting...my life An Acting Resource Page - An acting resource page that contains lots of auditions, a chat room, a message board, lots of links, and many more acting resources.
- Actors Cafe - Actor Resource: many links to other actor sites. Add your URL for free. Professional Actor Mailing list.
- Actor's Corner - An award-winning site from an ingenue actress that includes hard to reach industry/acting links.
- Actor's guide to jobs in NYC - Job listings for anyone interested in a career in acting in New York City.

Learning to summarize

- Collect all *1,551,990* summaries and associated web pages.
- From the collection of *(d,g)* pairs, learn a mapping $d \rightarrow g$.



The mapping will be probabilistic

Three models for gisting

- Bag of words
- Synonym-aware
- Readable

Naïve Bayes gisting

Let $\lambda(\bullet \mid d)$ denote the word distribution for web page d.

1. Select a length n for gist: $n \sim \phi$
2. For $i=1$ to n:
   - Pick a word from document: $w \sim \lambda(\bullet \mid d)$
   - Set $g_i \leftarrow w$

Naïve Bayes gisting

From a document d, the probability that a person will produce the summary $g=\{g_1\, g_2 \ldots\, g_n\}$ is

$$p(g \mid d) = \phi(n) \prod_{i=1}^{n} \lambda(g_i \mid d)$$

No "learning" involved

Don't even *need* OpenDirectory dataset!

This model is broken

cf: Ponte & Croft, 1998

Synonym-aware gisting

After drawing a word from web page, replace it with a synonym.

1. Select a length n for gist: $n \sim \phi$
2. For $i=1$ to n:
   - Pick a word from document: $w \sim \lambda(\bullet \mid d)$
   - Pick a synonym for w: $v \sim \sigma(\bullet \mid w)$
   - Set $g_i \longleftarrow v$

Synonym-aware gisting

From a document d, the probability that a person will produce the summary g is

$$p(g \mid d) \propto p(g)\, p(d \mid g)$$

(Take $p(g)$ uniform)

$$\propto \phi(n) \prod_{j=1}^{m} p(d_j \mid g)$$

cf: Ponte & Croft, 1998

σ: A statistical synonymy model

$\sigma(u \mid v)$ = probability that word v maps to word u

N: # of "known" words

The $\sigma(\bullet \mid w)$ model is a N by N stochastic matrix

Learn σ using methods from *statistical translation*.

EM training

Using 103,064 page/summary pairs

(24,231,164 words in web pages; 1,922,393 words in summaries)

```
job:  job 0.194   jobs 0.098 career 0.028 employment 0.028
camera: camera 0.137 cameras 0.045 photo 0.020 photography 0.014
IBM: UBM 0.130 business 0.085 solutions 0.019 support 0.017
ranches: ranches 0.089 springs 0.034 colorado 0.032 ranch 0.030
quilts: quilts 0.141 quilt 0.074 I 0.036 quilting 0.034
```


How to learn $\sigma(u \mid v)$?

Maximum-likelihood estimation. Given a collection $C=\{(d_1, g_1), (d_2, g_2) \ldots (d_N, g_N)\}$, find the σ for which $\Pi\, p(d_i \mid g_i)$ is maximal.

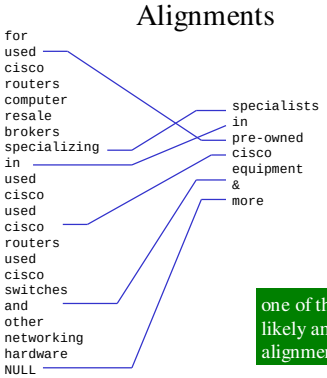
$$p(d \mid g) = \sum_a p(d, a \mid g)$$

$$= \sum_a p(a \mid g)\, p(d \mid a, g)$$

$$= \sum_a p(a \mid g) \prod_{i=1}^{m} \sigma(d_i \mid g_{a_i})$$

a: an "alignment" between d and g

Alignments



one of the more likely among 22^7 alignments

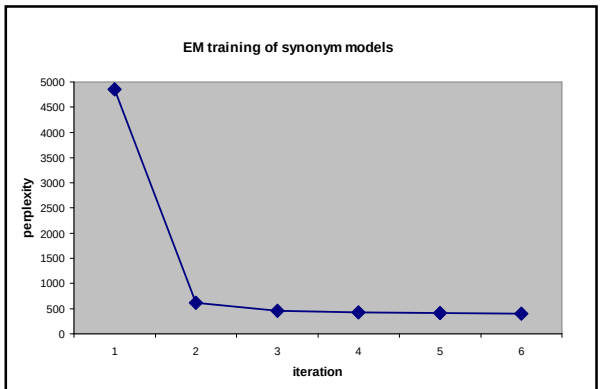
EM training

- Start with uniform $\sigma(u \mid v)$ model
- (E-step): accumulate (u,v) counts:

$$c(u,v) = \sum_{d,g} \sum_{a} p(d,a \mid g) \sum_{j=1}^{m} \delta(u,d_j)\delta(v,g_{a_j})$$

- (M-step): normalize counts

$$\sigma(u \mid v) = \frac{c(u,v)}{\sum_{u} c(u,v)}$$



Readable gisting

Previous approaches don't even *attempt* to produce a coherent string of words.

1. Select a length n for gist: $n \sim \phi$
2. Find, by searching, the gist g which maximizes $p(d \mid g)p(g)$

"fidelity" **"coherence"**

Language models

$$p(g_1 g_2 \ldots g_n) = \prod_{i=1}^{n} p(g_i \mid g_1 g_2 \ldots g_{i-1})$$

A good model will assign a higher probability to more likely sequences of words:

p("The best way to cook a rabbit") >
p("cook best the way rabbit to a")

A simple but powerful model:

$$p(g_i \mid g_1 g_2 \ldots g_{i-1}) \approx p(g_i \mid g_{i-2} g_{i-1})$$

Evaluation

Could use:

- word overlap between correct and hypothesized summaries
- user-studies (reading-comprehension questions)
- qualitative

Law Offices of Anahit N. Akbarian
100 N. Brand Blvd., Suite 200
Glendale, , CA 91207; USA
(818)247-9322
Pinnkee@aol.com

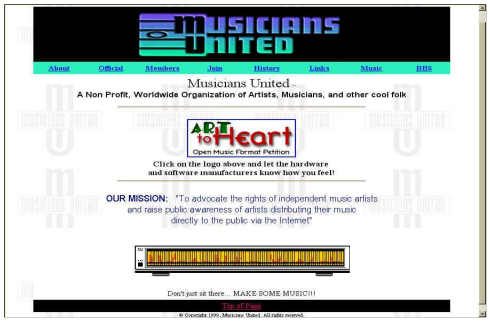
The Law Offices of Analit N. Akbarian provides legal services in the following areas of law: - Immigration -Visa -Green Cards -Citizenships -Work Authorizations -Deportation - Workers' compensation - Drunk driving - DMV Hearings - Personal injury -Attorney available for court appearances.

Home Page | Contact | Practice Areas | Links | Disclaimer

Firms Online • FindLaw • LegalMinds • LegalNews
FindLaw Firms Online -- Web Sites for the Legal Community.
Copyright 1998. FindLaw Disclaimer

OpenDirectory Gist: immigration visa green cards citizenships work authorizations deportation workers compensation drunk driving dmv hearings personal injury

Ocelot Gist: services for bankrupty and criminal



OpenDirectory Gist: to advocate the rights of independent music artists and raise public awareness of artists distributing their music directly to the public via the internet

Ocelot Gist: the music business and industry artists raise awareness rock and jazz

Where to go from here?

- Use layout information as a guide
- OCR for inlined images
- More intelligent language models
- Translingual summarization

Carnegie Mellon
School of Computer Science

Thanks...

Defense Advanced Research Projects Agency DARPA

IBM

Claritech Corporation

Carnegie Mellon
School of Computer Science