The Web Economy: Technical Challenges and Research Opportunities

Berthier Ribeiro-Neto
Google Engineering, Belo Horizonte
Federal University of Minas Gerais

March 2008

Summary

- The e-Publishing Age
- The Web Economy
- The Problem
- Solution 1: Syntactical Matching
- Solution 2: A GP Framework
- Conclusions

The e-Publishing Age

- The Web has changed the way people look for information, pay their bills, buy their goods, search for a new apartment. And this is just the beginning … after all, the Web was created in 1991.
- Why is the Web such a success? What is the single most important characteristic of the Web that makes it so revolutionary?
- Freedom to publish!

The e-Publishing Age

She finished the first draft of her novel in 1796. The first attempt of publication was refused without a reading. The novel was only published 15 years later!

She got a flat fee of $110, which meant that she was not paid anything for the many subsequent editions. Further, her authorship was anonymized under the reference "By a Lady".

The e-Publishing Age

"Pride and Prejudice" is the second best loved novel in the UK ever, after "The Lord of the Rings"!

It has been the subject of six TV series and five film versions. The last of these, starring Keira Knightley and Matthew Macfadyen, grossed over 100 million dollars.

Jane Austen published anonymously her entire life. Throughout the 20th century, her novels have never been out of print.

The e-Publishing Age

Jane Austen was discriminated because there was no freedom to publish in the beginning of the 19th century.

The Web, unleashed by the inventiveness of Tim Berners-Lee, changed this once and for all.

The Web moved mankind into a new era, into a new time, into the e-Publishing Age!

The e-Publishing Age

- The e-Publishing Age is characterized by a paradigm shift.
- Before, publishing was done by a select group of people, journalists, editors, authors, who either worked for major media companies or entered into agreement with them. In that world, all power on what gets published sits with the publishing companies.
- In the e-Publishing Age, dozens of millions of people became authors, journalists, editors. As a result, publishing by the masses became the rule of thumb. There is no centralized control anymore; tremendous power has been unleashed!

The e-Publishing Age

- Publishing by the masses:
  - MySpace, Orkut
  - YouTube
  - Flickr
  - Wikipedia
- In the e-Publishing Age, we leave the world of closed information gardens and enter into the world of one, huge, free, open information garden!

Summary

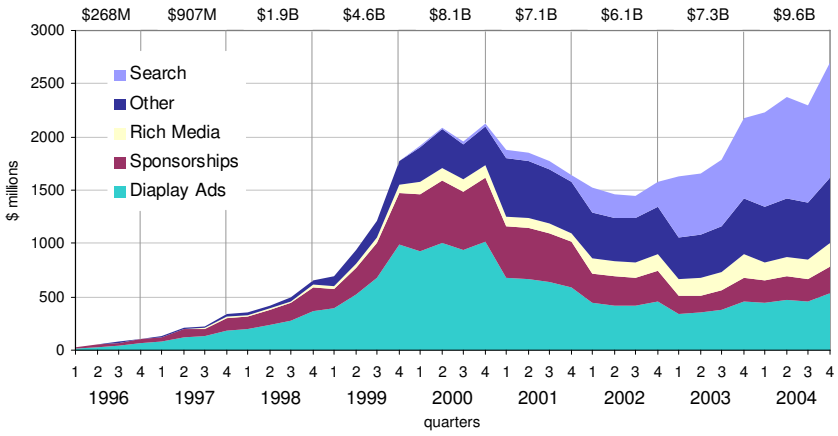
- The e-Publishing Age
- The Web Economy
- The Problem
- Solution 1: Syntactical Matching
- Solution 2: A GP Framework
- Conclusions

The Web Economy

- The open information garden counts with a viable business model!
- Worldwide advertising is expected to grow at a 6.3% annual rate from $383 billion in 2005 to $521 billion in 2010.
- In these 5 years, Web Advertising is expected to grow at a 18% rate reaching $51 billion in 2010 (Source: PricewaterhouseCoopers)
  - The current boom of web advertising is related to the success of a particular kind of advertising, Search Advertising

The Web Economy

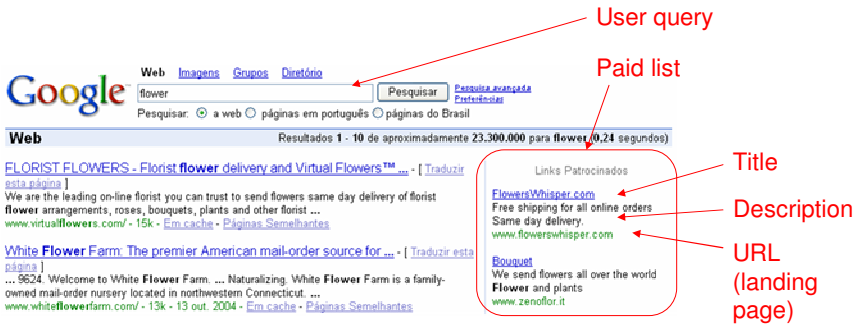
Web Advertising Revenue Growth Comparisons, 1996-2004, USA (Source: IAB)



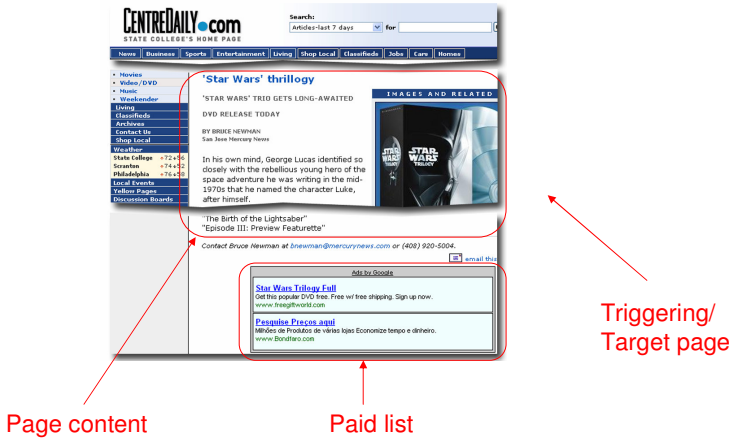
The Web Economy

- The Web economy is all about advertising.
- In Web advertising, an advertiser is given prominent position according to:
  - Relevance of its advertisements (ads) to the users
  - Amount it is willing to pay for a click on an ad (pay per performance)
- Main Categories
  - Keyword-targeted advertising (KTA)
  - Content-targeted advertising (CTA)

Keyword-targeted Advertising (KTA)
The Golden Search



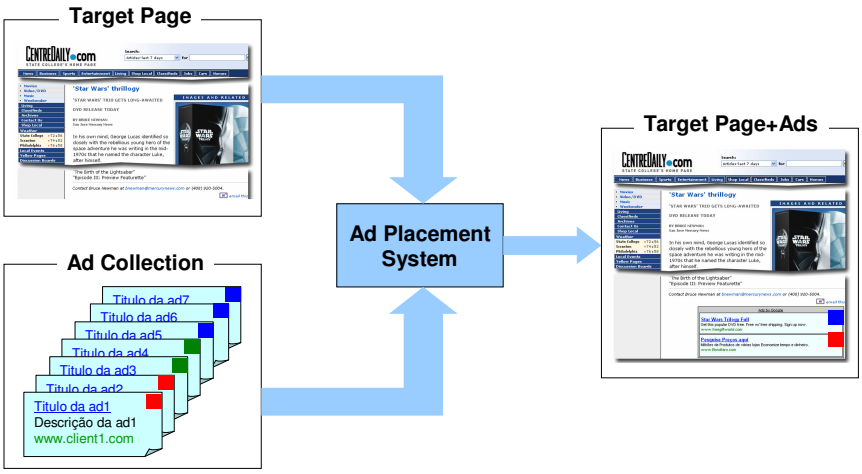
Content-targeted Advertising (CTA)



Summary

- The e-Publishing Age
- The Web Economy
- The Problem
- Solution 1: Syntactical Matching
- Solution 2: A GP Framework
- Conclusions

The Problem: Content-targeted Advertising (CTA)

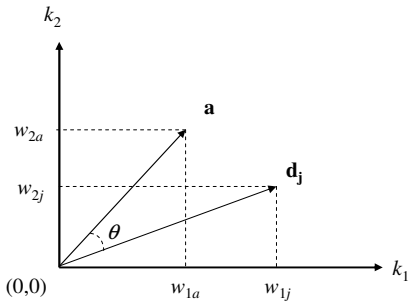


Summary

- The e-Publishing Age
- The Web Economy
- The Problem
- Solution 1: Syntactical Matching
- Solution 2: A GP Framework
- Conclusions

Solution 1: Syntactical Matching Algorithms

- Simple strategy consists in matching the ad to the page contents using, for instance, the **Vector Space Model**



k_1 and k_2 are terms

$$\mathbf{a} = (w_{1a}, w_{2a})$$
$$\mathbf{d}_j = (w_{1j}, w_{2j})$$

$$ssim(\mathbf{a}, \mathbf{d}_j) = \cos\theta$$

Experiments with Syntactical Matching Strategies

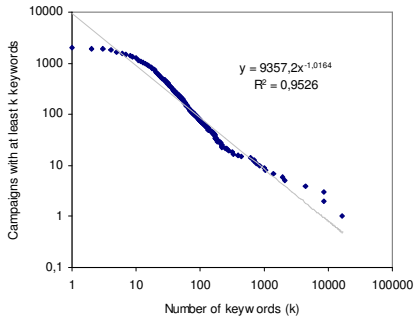
- Ads Collection
  - 1,744 advertisers
  - 93,972 ads in 2,029 campaigns
  - 68,238 keywords
- Test Collection
  - 100 pages of a Brazilian newspaper
  - Topics include economy, sports, culture, politics etc.
- Match ads to the 100 Web test pages.

Solution 1
A Comparison of Our 5 Simple Strategies

Methods	Precision			PAVG@3		PAVG	
	@1	@2	@3	Score	Gain (%)	Score	Gain (%)
AD	0.410	0.365	0.287	0.257	-	0.110	-
AD_KW	0.510	0.395	0.320	0.296	+15.2	0.124	+12.7
KW	0.460	0.395	0.353	0.323	+25.7	0.136	+23.6
ANDKW	0.490	0.425	0.400	0.364	+41.6	0.160	+45.5
AAK	0.510	0.495	0.460	0.412	+60.3	0.175	+59.1

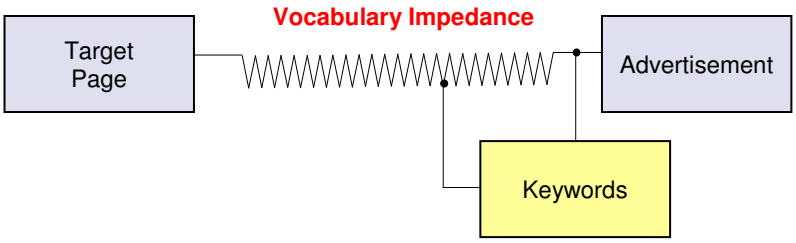
An Observation: The Vocabulary Mismatch

- Vocabulary mismatch between the ads and the target page.

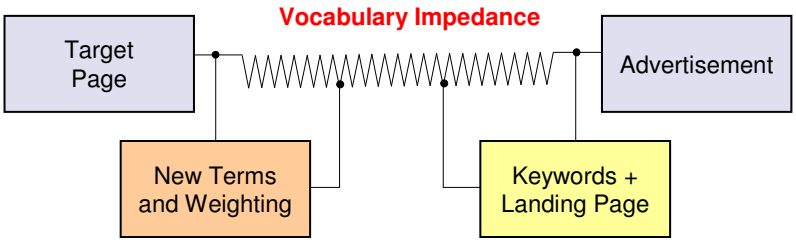


- Content of the target page might be diffuse
  - Because it includes multiple topics
  - Also, a non-central topic might offer a good opportunity for advertising

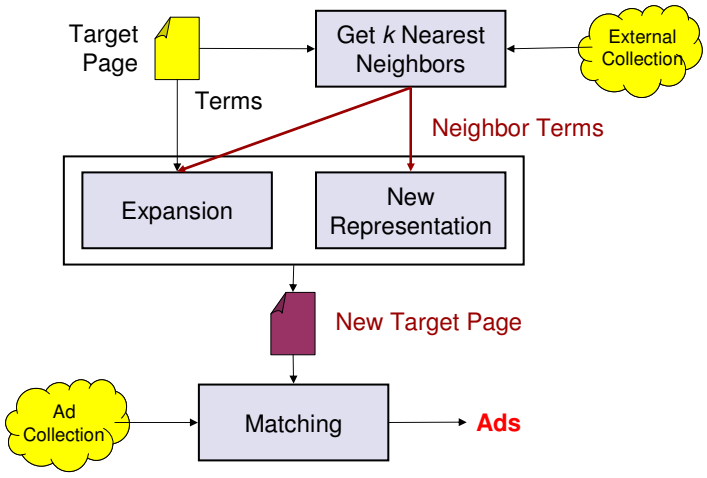
Solution 1: Impedance Coupling -- An Advanced Approach for Dealing with the Vocabulary Mismatch



Solution 1: Impedance Coupling -- An Advanced Approach for Dealing with the Vocabulary Mismatch



Impedance Coupling: Obtaining New Terms to Add to the Target Page



Comparing the Original and the Expanded Target Pages

- Argentinean wines produced with grapes from Bordeaux, France

Rank	Original Target Page		New Target Page	
	term	score	term	score
1	argentina	0.090	wines	0.251
2	obtained	0.047	wine	0.140
3	class	0.036	whites	0.091
4	whites	0.035	red	0.057
5	french	0.031	grape	0.051
6	origin	0.029	bordeaux	0.045
7	france	0.029	acideness	0.038
...	...	...	...	...
35	wines	0.010	-	-
...	...			

Terms in red are not common

Solution 1: A Comparison of Advanced Syntatic Matching Methods

Methods	Precision			PAVG@3		PAVG	
	@1	@2	@3	Score	Gain (%)	Score	Gain (%)
AAK	0.510	0.495	0.460	0.412	-	0.175	-
AAK_H	0.510	0.510	0.463	0.421	+ 2.2*	0.181	+ 3.4*
AAK_T	0.663	0.582	0.534	0.498	+ 20.9	0.231	+ 32.0
AAK_EXP	0.700	0.610	0.583	0.554	+ 34.5	0.248	+ 41.7
AAK_EXP_H	0.690	0.615	0.570	0.533	+ 29.4	0.250	+ 42.9

(*) indicates not significant gain

Solution 1: Syntactical Matching

- Keywords are essential to improve ad placement precision. Enforcing their appearance led to the best method among the simple methods.
- Further improvement was possible by expanding the target page with terms gathered from pages of similar content, with the purpose of reducing vocabulary mismatch.
- Berthier Ribeiro-Neto, Marco Cristo, Paulo Golgher, Edleno Moura. *Impedance coupling in content-targeted advertising*. ACM SIGIR 2005, Salvador, pages 496-503.

Summary

- The e-Publishing Age
- The Web Economy
- The Problem
- Solution 1: Syntactical Matching
- Solution 2: A GP Framework
- Conclusions

Solution 3: Genetic Programming (GP)

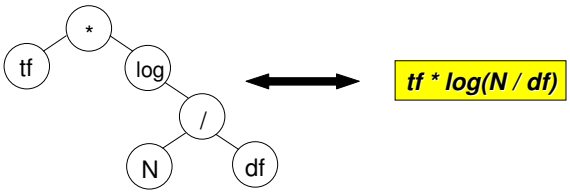
- Use Genetic Programming (GP) to evolve the best functions for ranking ads.
- Why GP?
  - GP is able to learn intrinsic characteristics of Content-Targeted Advertising.
  - GP can deal with multi-objective functions.

Solution 3: Genetic Programming (GP)

- GP is a machine learning technique, inspired by biological evolution, that we use to find optimized ranking functions.
- Components
  - Individuals: ranking functions for ads
  - Fitness function: to estimate the ranking quality
- At training time, *evolve individuals* in the search of better ads ranking functions.
  - Evolution requires applying genetic operators: *reproduction, mutation, crossover*.

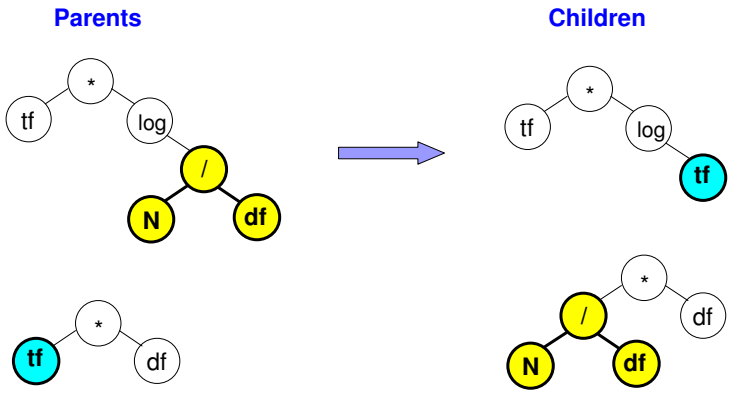
Solution 3: Genetic Programming

- Example of a ranking function



Solution 3: Evolution (applying genetic operators)

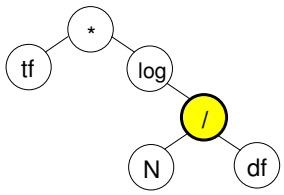
- Cross-over



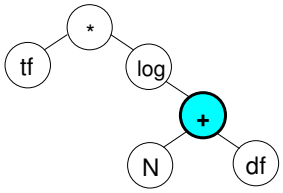
Solution 3: Evolution (applying genetic operators)

- Mutation
  - A randomly selected tree is replaced by a new subtree also created randomly.

Original individual



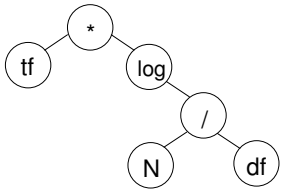
Mutated individual



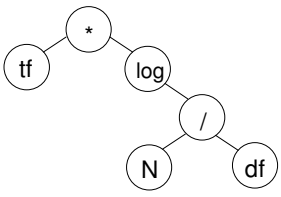
Solution 3: Evolution (applying genetic operators)

- Reproduction
  - Copy the selected individual program to the new population.

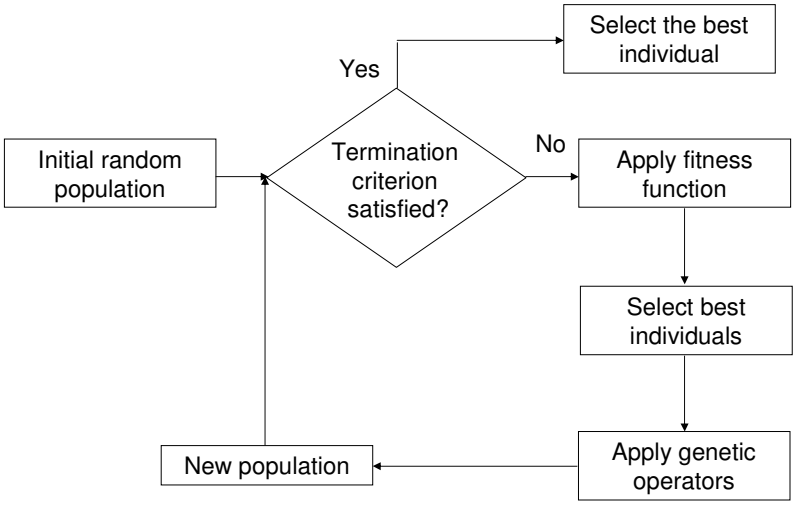
Individual



Individual



Solution 3: GP overview



Solution 3: Experimental Design

- Random sampling
  - Training set: 50 pages
  - Validation set: 30 pages
  - Test set: 20 pages
  - This is not practical in the Web!?

Solution 3: Exactly three ads per page

Methods	Hits/Suggestions				pavg@3	
	#1	#2	#3	Total	Score	Gain
Baseline	9/20	5/20	9/20	23/60	0.314	-
GP	**14/20**	11/20	7/20	32/60	**0.508**	**+61.7%**

- Anisio Lacerda, Marco Cristo, Marcos Goncalves, Weiguo Fan, Nivio Ziviani, Berthier Ribeiro-Neto. *Learning to Advertise.* ACM SIGIR 2006.

Conclusions on Content-targeted Advertising (CTA)

- Solution 1: careful consideration of the available evidence leads to enhancements in CTA, using syntactical matching.
  - Reduce the vocabulary mismatch (between ads and pages) by expanding the content of the pages.
- Solution 2: Genetic programming can be used to learn advanced ranking functions for ads.
  - Needs training set for learning parameters. How large?

The e-Publishing Age is here, join in!

Thank you!