

LaSEWeb: Automating Search Strategies over Semi-Structured Web Data

Oleksandr Polozov

University of Washington
polozov@cs.washington.edu

Sumit Gulwani

Microsoft Research
sumitg@microsoft.com



Microsoft®
Research

Motivation: search engine micro-segments

The image shows a Bing search result for the query "Bubble gum inventor". The search bar at the top contains the text "Bubble gum inventor" and a magnifying glass icon. Below the search bar, it indicates "MS Beta", "409,000 RESULTS", and "Any time". The main result is for "Bubble gum inventor" with the name "Walter Diemer" in large blue text and "Data from Freebase" below it. To the left, there are three search results: "Bubble Gum History - Invention of Bubble Gum" from www.idealfinder.com, "4. Who invented bubble gum?" from whoinventedthe.com, and "Images of Bubble gum inventor" from bing.com/images. To the right, there is a detailed knowledge panel for "Walter Diemer" with a biography, birth and death dates, spouse information, and inventions.

bing Bubble gum inventor

MS Beta 409,000 RESULTS Any time ▾

Bubble gum inventor
Walter Diemer
Data from Freebase

[Bubble Gum History - Invention of Bubble Gum](#)
www.idealfinder.com/history/inventions/bubblegum.htm ▾
Fascinating facts about the invention of Bubble Gum by Walter Diemer in 1928.

[4. Who invented bubble gum ?](#)
whoinventedthe.com/bubble-gum ▾
THE HISTORY OF BUBBLEGUM Like a lot of inventions, Watler Diemer (1904-1998) didn't get it right on the first attempt. In 1928, **Walter Diemer** who worked as an ...

[Images of Bubble gum inventor](#)
bing.com/images

Walter Diemer

Walter E. Diemer was an accountant and inventor of bubble gum. Born and raised in Philadelphia, Pennsylvania, Diemer was working as an accountant at Fler in 1926 when the company president sought to cut costs by making their own gum base. The company's founder, Frank Henry Fler, had previously made a batch of bubble gum in 1906 ... +

en.wikipedia.org

Born: Jan 08, 1904 · [Philadelphia, Pennsylvania](#)

Died: Jan 09, 1998

Spouse: [Florence Diemer](#) · [Adelaide Diemer](#)

Inventions: [Bubble gum](#)

Timeline

1926: Born and raised in Philadelphia, Pennsylvania, Diemer was

Motivation: search engine micro-segments

The image shows a Bing search results page for the query 'radio inventor'. The search bar at the top left contains the text 'radio inventor' and a magnifying glass icon. Below the search bar, there are suggestions: 'Also try: Who Invented The Radio · The History of the Radio · Facts About The Fi...'. The search results are displayed as follows:

- 13,200,000 RESULTS** Any time ▾
- Invention of radio - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/Invention_of_radio ▾
Many people were involved in the invention of radio in its current form. Experimental work on the connection between electricity and magnetism began around ...
[Wireless signalling ...](#) · [Theory of ...](#) · [Maxwell and the ...](#)
- The Invention of Radio - About.com Inventors**
inventors.about.com > ... > [Famous Inventors - M](#) ▾
An outline of the development of radio technology - radio owes its development to two other inventions the telegraph and the telephone.
- Related searches for radio inventor**
[Who Invented The Radio](#) [Airplane Inventor](#)
[The History of the Radio](#) [First Radio](#)
[Facts About The First Radio](#) [When was the First Radio Broadcast](#)
- Inventor of the Radio - edHelper.com**
www.edhelper.com/ReadingComprehension_33_61.html ▾
1 Who invented the telephone? Easy -- Alexander Graham Bell. Who invented the airplane? Easy - Orville and Wilbur Wright. Who invented the radio?

On the right side of the page, there is a section titled 'See results for' with a small image of a radio and the text: 'Radio Radio is the radiation of electromagnetic signals through the atmosphere or free space. Information, such as sound, is carried by systematically changing some pro...'. Below this, there is a section titled 'Ads related to radio inventor' with two advertisements:

- Radio Invention at Amazon**
www.amazon.com/Radio Invention
Free 2-Day Shipping w/ Amazon Prime! Low Prices on Millions of Books.
Departments: Education & Reference, Science & Math and more
amazon.com is rated ★★★★★ on Bing (6134 reviews)
- Inventors Radio**
About.com/Inventors Radio
Search for Inventors Radio. Find Expert Advice on About.com.
[See your ad here »](#)

At the bottom right, there is a section titled 'Related searches'.

Motivation: search engine micro-segments

MS Beta 12,800,000 RESULTS Any time ▾

The Hobbit: The Battle of the Five Armies release date
December 17, 2014
[Data from IMDb](#)

['The Hobbit' Part 3 Release Date: 'Battle Of The ...](#)
www.idigitaltimes.com/.../hobbit-part-3-release-date-title-change.htm ▾
May 11, 2014 · Now, I'll admit right up front: "The Hobbit" part 3 release date is hardly the beginning of my discontent with Peter Jackson's prequel trilogy.

['The Hobbit' Part 3 Release Date: 'There And Back Again ...](#)
www.idigitaltimes.com/.../the-hobbit-part-3-release-date-2014.htm ▾
Jan 20, 2014 · So I'm finally going to see The Hobbit Part 2, because I'm a sucker, and it got my mind-grapes wondering: when is "The Hobbit" Part 3 release date?

[News about Hobbit 3 Release Date](#)
bing.com/news

 [The Hobbit 3 trailer. release date announced on Facebook](#)
[Christian Today](#) · 8 hours ago
"One day, they'll remember - remember everything that happened. The good, the bad, those who survived and those that did not." That was what Bilbo Baggins said...

['Hobbit' 3 Release Date: Australia To Wait Longer Than New Zealand and Others](#)
[International Business...](#) · 6 days ago

[New Line reveals The Hobbit: The Desolation of Smaug Extended Editions on Blu-ray](#)
[hd-report.com](#) · 12 days ago

★ [Stay up-to-date on this news topic](#)

The Hobbit: The Battle of the Five Armies (2014)


en.wikipedia.org

The Hobbit: The Battle of the Five Armies is an upcoming 2014 epic fantasy adventure film, directed by Peter Jackson and written by Jackson, Fran Walsh, Philippa Boyens, and Guillermo del Toro. It will be the third and final installment in the three-part film adaptation based on The Hobbit by J. R. R. Tolkien. It is preceded by An Unexpected Journ... +
en.wikipedia.org

Summary: NR · Fantasy
Release date: Dec 17, 2014
Director: [Peter Jackson](#)
Prequel: [The Hobbit: The Desolation of Smaug](#)
Screenwriters: [Peter Jackson](#) · [Guillermo del Toro](#) · [Fran Walsh](#) · [Philippa Boyens](#)
Story by: [J. R. R. Tolkien](#)

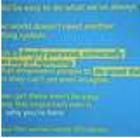
Motivation: search engine micro-segments

The image shows a Bing search results page for the query "windows 9 release date". The search bar at the top left contains the text "windows 9 release date" and a magnifying glass icon. Below the search bar, the Bing logo is visible on the left, and the search results are displayed. The results include several news articles with titles and snippets, a "Related searches" section on the right, and an advertisement for Microsoft's online store at the bottom right.

bing windows 9 release date

MS Beta 57,600,000 RESULTS Any time ▾

[News about Windows 9 Release Date](#)
[bing.com/news](#)

 [Windows 9 release date, beta, preview, start menu & screenshots](#)
TrustedReviews · 7 hours ago
When is the **Windows 9** release date and how will **Windows 9** differ from **Windows 8**? Read on as we explore potential **Windows 9** release dates, leaked **Windows 9**...

[Windows 9 release date, price, features: 30 September marked for unveiling](#)
PC Advisor · 8 hours ago

[Windows 9 Unveiling Expected September 30](#)
Value Walk · 9 hours ago

See also: [More stories](#) · [Top stories](#)

★ [Stay up-to-date on this news topic](#)

[Windows 9 release date, news and rumors | News | TechRadar](#)
[www.techradar.com/.../windows-9-release-date-news-and-rumours-1029245](#) ▾
Windows 9 release date, news and rumors - What will we see in the next complete version of **Windows**? And when will we see it? Buying advice from the leading ...

Related searches

- [Windows 9 Beta](#)
- [Windows 9 2014](#)
- [Windows 9 Download](#)
- [Download Windows 9 Beta](#)
- [Microsoft Windows 9](#)
- [Windows 8 to Windows 9](#)
- [When is Windows 9 Coming](#)
- [Microsoft Windows 9 Download](#)

Ads related to windows 9 release date

[Buy from Microsoft® Store](#) 
[MicrosoftStore.com/US_Online_Store](#)
 **FREE (855) 321-3207**
Official Site. Download Software. Buy PCs, Xbox, Kinect, & Surface!
Categories: Surface, Xbox, Computers, Windows Phone, Windows and more

Repetitive search tasks

Structured databases

- Precise, but limited in content
- No time-sensitive information
- Provide no context (sources)

Repetitive search tasks

Structured databases

- Precise, but limited in content
- No time-sensitive information
- Provide no context (sources)



Web mining scripts

- Two extremes:
 - Powerful ML, which has to be re-learned for each micro-segment
 - Fragile HTML layout parser
- Inaccessible for end-users

LaSEWeb Query Language

- A semantic scripting language for semi-structural information extraction from the Web
- Models natural patterns from the humans' search strategies

LaSEWeb interpreter

- Explores multiple webpages, clusters different answer candidates, and provides context for each answer
- Makes use of state-of-the-art NLP/ML/PL algorithms

Example: phone number



Sumit Gulwani

Primary Affiliation: Senior Researcher @ [Microsoft Research](#) (Redmond Lab)

Other Affiliations: Adjunct Faculty @ [IIT Kanpur](#)

Affiliate Faculty @ [Univ. of Washington](#)

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052

Email: sumitg at microsoft dot com

Phone: (425) 706-7709

Fax: (425) 936 7329 (attn: sumitg)

\vec{v} = (“Sumit Gulwani”)

let η_t = *Emphasized*(v_1) **in**

let η_b = *AttributeLookup*(*Syn*(“phone”), ℓ_a) **in**
Union(η_t, η_b)

where *Regex*(ℓ_a , “\(\d+\)\W*\d+\W*\d+”)

where *Layout*(η_t, η_b , Down) **and** *Nearby*(η_t, η_b)

Example: phone number



Sumit Gulwani

Primary Affiliation: Senior Researcher @ [Microsoft Research](#) (Redmond Lab)

Other Affiliations: Adjunct Faculty @ [IIT Kanpur](#)

Affiliate Faculty @ [Univ. of Washington](#)

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052

Email: sumitg at microsoft dot com

Phone: (425) 706-7709

Fax: (425) 936 7329 (attn: sumitg)

$\vec{v} = (\text{“Sumit Gulwani”})$

let $\eta_t = \textit{Emphasized}(v_1)$ **in**

let $\eta_b = \textit{AttributeLookup}(\textit{Syn}(\text{“phone”}), \ell_a)$ **in**
 $\textit{Union}(\eta_t, \eta_b)$

where $\textit{Regex}(\ell_a, "\d+\W*\d+\W*\d+")$

where $\textit{Layout}(\eta_t, \eta_b, \text{Down})$ **and** $\textit{Nearby}(\eta_t, \eta_b)$

- Visual attributes

Example: phone number



Sumit Gulwani

Primary Affiliation: Senior Researcher @ [Microsoft Research](#) (Redmond Lab)

Other Affiliations: Adjunct Faculty @ [IIT Kanpur](#)

Affiliate Faculty @ [Univ. of Washington](#)

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052

Email: sumitg at microsoft dot com

Phone: (425) 706-7709

Fax: (425) 936 7329 (attn: sumitg)

\vec{v} = (“Sumit Gulwani”)

let η_t = *Emphasized*(v_1) **in**

let η_b = *AttributeLookup*(*Syn*(“phone”), ℓ_a) **in**

Union(η_t, η_b)

where *Regex*(ℓ_a , “\(\d+\)\W*\d+\W*\d+”)

where *Layout*(η_t, η_b , Down) **and** *Nearby*(η_t, η_b)

- Visual attributes
- Implicit table detection

Example: phone number



Sumit Gulwani

Primary Affiliation: Senior Researcher @ [Microsoft Research](#) (Redmond Lab)

Other Affiliations: Adjunct Faculty @ [IIT Kanpur](#)

Affiliate Faculty @ [Univ. of Washington](#)

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052

Email: sumitg at microsoft dot com

Phone (425) 706-7709

Fax: (425) 936 7329 (attn: sumitg)

\vec{v} = (“Sumit Gulwani”)

let η_t = *Emphasized*(v_1) **in**

let η_b = *AttributeLookup*(*Syn*(“phone”), ℓ_a) **in**
Union(η_t, η_b)

where *Regex*(ℓ_a , “\(\d+\)\W*\d+\W*\d+”)

where *Layout*(η_t, η_b , Down) **and** *Nearby*(η_t, η_b)

- Visual attributes
- Implicit table detection
- Linguistic patterns

Example: phone number



Sumit Gulwani

Primary Affiliation: Senior Researcher @ [Microsoft Research](#) (Redmond Lab)

Other Affiliations: Adjunct Faculty @ [IIT Kanpur](#)

Affiliate Faculty @ [Univ. of Washington](#)

Address: Microsoft Corporation, One Microsoft Way, Redmond, WA, 98052

Email: [sumitg at microsoft dot com](mailto:sumitg@microsoft.com)

Phone: (425) 706-7709

Fax: (425) 936 7329 (attn: sumitg)

$\vec{v} = (\text{"Sumit Gulwani"})$

let $\eta_t = \textit{Emphasized}(v_1)$ **in**

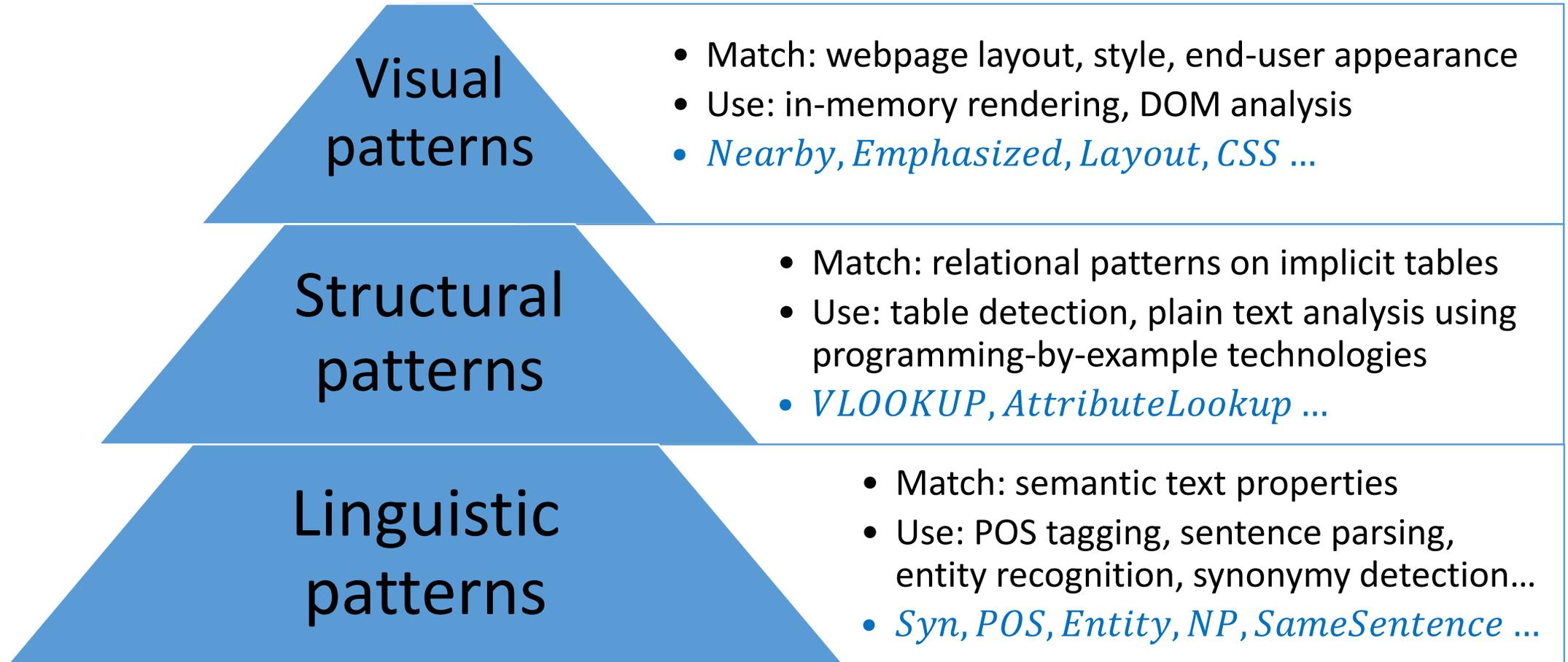
let $\eta_b = \textit{AttributeLookup}(\textit{Syn}(\text{"phone"}), \ell_a)$ **in**
 $\textit{Union}(\eta_t, \eta_b)$

where $\textit{Regex}(\ell_a, "\d+\W*\d+\W*\d+")$

where $\textit{Layout}(\eta_t, \eta_b, \text{Down})$ **and** $\textit{Nearby}(\eta_t, \eta_b)$

- Visual attributes
- Implicit table detection
- Linguistic patterns
- Clustering across webpages

Language Structure



[1] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In ACL, 2005.

[2] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In ACL, 2003.

[3] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In HLT-NAACL, 2003.

[4] C. Quirk, P. Choudhury, J. Gao, H. Suzuki, K. Toutanova, M. Gamon, W.-t. Yih, L. Vanderwende, and C. Cherry. MSR SPLAT, a language analysis toolkit. In ACL, 2012.

[5] W.-t. Yih, G. Zweig, and J. C. Platt. Polarity inducing latent semantic analysis. In ACL, 2012.

[6] S. Gulwani. Automating string processing in spreadsheets using input-output examples. In POPL, 2011.

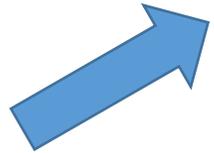
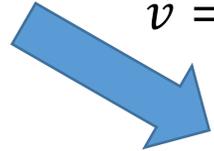
[7] M. J. Cafarella, A. Halevy, and J. Madhavan. Structured data on the web. In CACM 54.2 (2011): 72-79.

Program interpreter: “user emulation” algorithm

Program interpreter: “user emulation” algorithm



$\vec{v} = (\text{"computer"})$

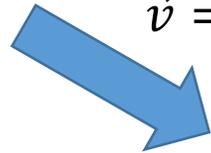


LaSEWeb
“inventors”
MS script

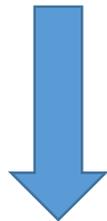
Program interpreter: “user emulation” algorithm



$\vec{v} = \text{"computer"}$



Seed query



[Who invented the computer - Answers.com](#)

[wiki.answers.com](#) > Categories > Technology > Inventions ▾

Inventor of the Computer Many say the first computer is the "difference engine." The first of these devices was conceived in 1786 by J.H. Müller.

[The History of Computers - Computer History Timeline](#)

[inventors.about.com/library/blcoindex.htm](#) ▾

A history of computers and the inventors involved with each computer invention milestone - a timeline with detailed history features.

[Computer - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Computer](#) ▾

A computer is a general purpose device that can be programmed to carry out a set of arithmetic or logical operations automatically. Since a sequence of operations can ...

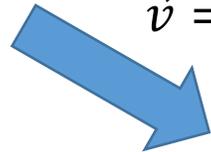
[Etymology](#) · [History](#) · [Programs](#) · [Components](#) · [Misconceptions](#) · [Further topics](#)



Program interpreter: “user emulation” algorithm



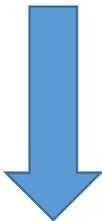
$\vec{v} = \text{"computer"}$



LaSEWeb Engine



Seed query



[Who invented the computer - Answers.com](#)
wiki.answers.com > Categories > Technology > Inventions ▼
Inventor of the Computer Many say the first computer is the "difference engine." The first of these devices was conceived in 1786 by J.H. Müller.

[The History of Computers - Computer History Timeline](#)
inventors.about.com/library/blcoindex.htm ▼
A history of computers and the inventors involved with each computer invention milestone - a timeline with detailed history features.

[Computer - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Computer ▼
A computer is a general purpose device that can be programmed to carry out a set of arithmetic or logical operations automatically. Since a sequence of operations can ...
Etymology · History · Programs · Components · Misconceptions · Further topics



“John Atanasoff”



“John Vincent Atanasoff”



“Charles Babbage”



“Babbage, C.”



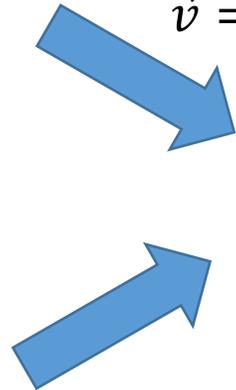
“konrad zuse”

LaSEWeb
“inventors”
MS script

Program interpreter: “user emulation” algorithm



$\vec{v} = (\text{"computer"})$



LaSEWeb Engine

Seed query



[Who invented the computer - Answers.com](#)

[wiki.answers.com > Categories > Technology > Inventions](#)

Inventor of the Computer Many say the first computer is the "difference engine." The first of these devices was conceived in 1786 by J.H. Müller.

[The History of Computers - Computer History Timeline](#)

[inventors.about.com/library/blcoindex.htm](#)

A history of computers and the inventors involved with each computer invention milestone - a timeline with detailed history features.

[Computer - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Computer](#)

A computer is a general purpose device that can be programmed to carry out a set of arithmetic or logical operations automatically. Since a sequence of operations can ...

[Etymology](#) · [History](#) · [Programs](#) · [Components](#) · [Misconceptions](#) · [Further topics](#)

$$\text{score}(C_i) = \frac{1}{|U|} \sum_{j=1}^{|U|} \sum_{s \in C_i} \frac{c(s, u_j)}{c(u_j)}$$



“John Atanasoff”

“John Vincent Atanasoff”

“Charles Babbage”

“Babbage, C.”

“konrad zuse”

Program interpreter: "user emulation" algorithm



\vec{v} = ("computer")

LaSEWeb Engine

Seed query



John Atanasoff (14.5%)

<http://www.computerhope.com>

<http://www.ehow.com>

<http://inventors.about.com>

Charles Babbage (10.5%)

<http://www.buzzle.com>

<http://www.ask.com>

...

[Who invented the computer - Answers.com](http://www.answers.com)

[wiki.answers.com](http://www.answers.com) > Categories > Technology > Inventions

Inventor of the Computer Many say the first computer is the "difference engine." The first of these devices was conceived in 1786 by J.H. Müller.

[The History of Computers - Computer History Timeline](http://www.inventors.about.com)

[inventors.about.com/library/blcoindex.htm](http://www.inventors.about.com)

A history of computers and the inventors involved with each computer invention milestone - a timeline with detailed history features.

[Computer - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Computer)

en.wikipedia.org/wiki/Computer

A computer is a general purpose device that can be programmed to carry out a set of arithmetic or logical operations automatically. Since a sequence of operations can ...

[Etymology](#) · [History](#) · [Programs](#) · [Components](#) · [Misconceptions](#) · [Further topics](#)

$$score(C_i) = \frac{1}{|U|} \sum_{j=1}^{|U|} \sum_{s \in C_i} \frac{c(s, u_j)}{c(u_j)}$$

"John Atanasoff"

"John Vincent Atanasoff"

"Charles Babbage"

"Babbage, C."

"konrad zuse"

LaSEWeb
"inventors"
MS script

Experiments

Micro-segment	# queries	Recall	Bing recall
ASCII code of a symbol	1,551	32.88%	0%
Calories in a food	9,207	71.80%	0%
Inventor of a product	8,994	75.91%	16.01%
Lyrics of a song	48,995	24.36%	0%
Phone number of a company	6,881	95.49%	0%
Population of a place	18,151	92.53%	57.58%
Release date of a product	12,339	97.24%	12.60%

Search task	Recall	Precision
Phone #	29/37	21/29
Affiliation	34/37	22/34
PhD institution	21/37	13/21
General chair	21/28	17/21
Invited talks	13/28	11/13
Average	71%	73%

- ~95% precision and 71% recall on factoid micro-segments
 - For micro-segments: Precision measured by random sampling, based on top-3 results
 - For end-user repetitive search tasks: Precision/recall measured manually
- Average execution time: ~5 sec/webpage
 - Depends on the rendering settings
- Current setting: offline deployment / database population

Summary & Future work

- Typical patterns of human search strategies in a scripting language for IE
 - Match semi-structured Web content
 - Existing cross-disciplinary technologies used as building blocks
 - Exploit information redundancy across multiple webpages
- Applications:
 1. Micro-segments of factoid questions in search engines
 2. Repeatable batch data extraction tasks for end-users
 3. Structured database population from free Web text
 4. English language comprehension problem generation
- Future work:
 - Automatic query execution plans in the language
 - Integration with “natural language → logic” engines

Summary & Future work

1. The principal characterized his pupils as _____ because they were pampered and spoiled by their indulgent parents.
2. The commentator characterized the electorate as _____ because it was unpredictable and given to constantly shifting moods.
(a) cosseted (b) disingenuous (c) corrosive (d) laconic (e) mercurial

2. Repeatable batch data extraction tasks for end-users
3. Structured database population from free Web text
4. English language comprehension problem generation

- Future work:
 - Automatic query execution plans in the language
 - Integration with “natural language → logic” engines

Summary & Future work

- Typical patterns of human search strategies in a scripting language for IE
 - Match semi-structured Web content
 - Existing cross-disciplinary technologies used as building blocks
 - Exploit information redundancy across multiple webpages
- Applications:
 1. Micro-segments of factoid questions in search engines
 2. Repeatable batch data extraction tasks for end-users
 3. Structured database population from free Web text
 4. English language comprehension problem generation
- Future work:
 - Automatic query execution plans in the language
 - Integration with “natural language → logic” engines

Thanks for listening!

Questions?