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

Motivation

A significant percent of search queries constitute repetitive tasks. Two most common examples are:

1. Batch data extraction, done by end-users.
2. Development of micro-segments of factoid question answering in search engines.

Typical solutions involve:

- Using a structured database (e.g. FreeBase) (limited in content; hard-coded; time-insensitive)
- Writing a data mining script (fragile; inapplicable for end-users)

Both solutions do not preserve any of the following end-users' search process patterns:

- Checking multiple webpages/answer candidates
- Exploring the context related to each answer
- Utilizing a semi-structured webpage format

Problem definition

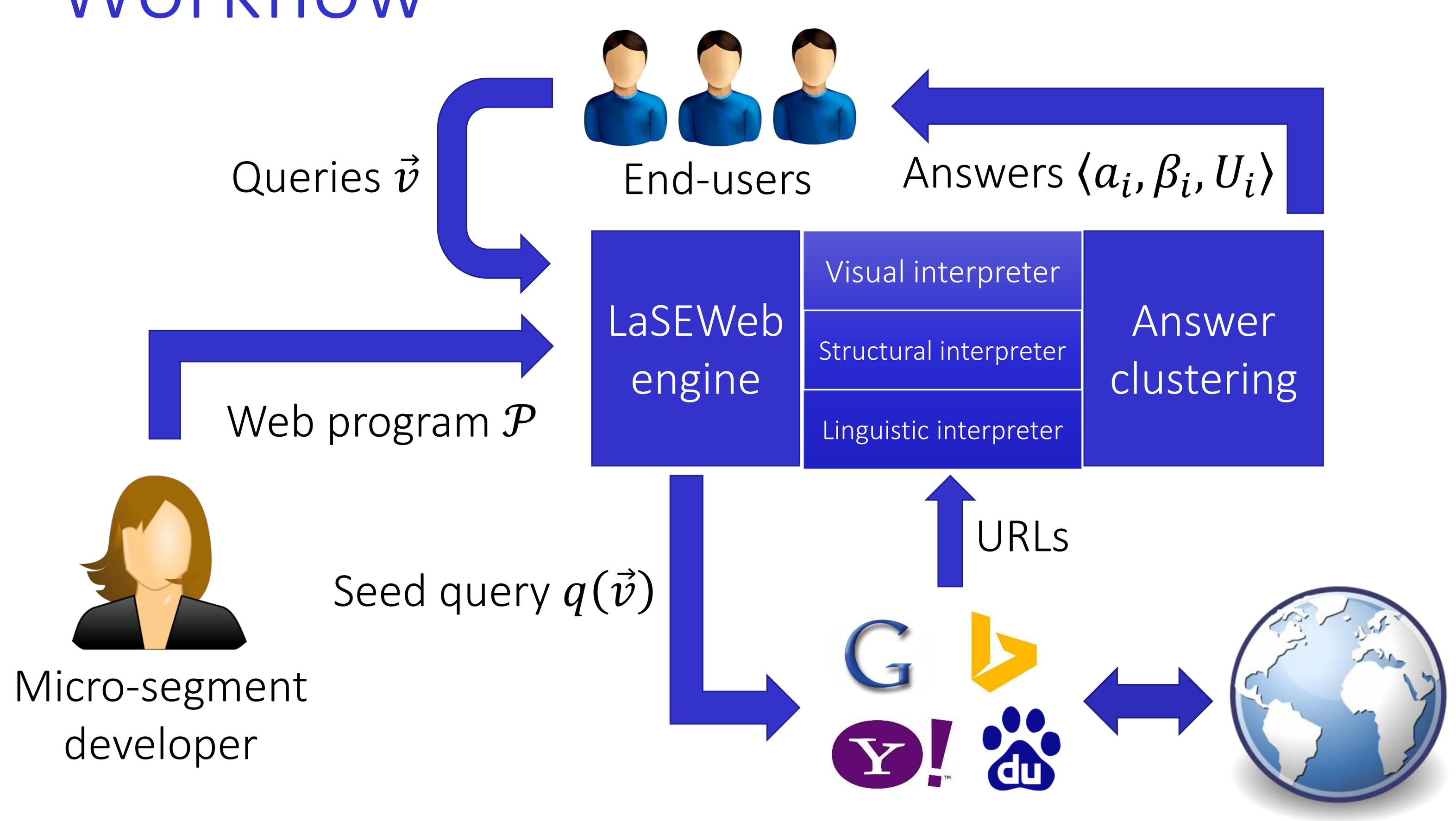
A **Web program $\mathcal{P}$** is a parameterized query that

- takes a tuple of user **query arguments $\vec{v}$**

and returns a set of:

- **answer strings a_i**
- ranked by their **confidence β_i**
- with a set of the corresponding **source URLs U_i**.

Workflow



LaSEWeb Query Language

A semantic scripting language for repetitive Web mining, based on the patterns in humans' search strategies. The set of patterns is modular, extensible, and is implemented using the state-of-the-art ML/NLP algorithms.

$$\text{LaSEWeb query } \mathcal{Q} := \mathsf{FW}(\mathcal{B}, \Psi) \mid \mathcal{Q}_1 \vee \mathcal{Q}_2$$

$$\text{Visual expression } \mathcal{B} := \mathcal{S} \mid \mathsf{Union}(\mathcal{B}_1, \mathcal{B}_2) \mid \eta \colon \mathcal{B}$$

$$\text{Visual constraint } \Psi := \mathsf{Nearby}(\eta_1, \eta_2) \mid \mathsf{Emphasized}(\eta) \mid \mathsf{Layout}(\eta_1, \eta_2, d) \mid \Psi_1 \wedge \Psi_2 \mid \Psi_1 \vee \Psi_2 \mid \neg\Psi \mid \mathsf{true} \mid \mathsf{false}$$

$$\text{Direction } d \in \{\mathsf{Up}, \mathsf{Down}, \mathsf{Left}, \mathsf{Right}\}$$

$$\text{Structural expression } \mathcal{S} := \mathcal{L} \mid \mathsf{VLOOKUP}(\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3) \mid \mathsf{AttrLookup}(\mathcal{L}_1, \mathcal{L}_2)$$

$$\text{Linguistic expression } \mathcal{L} := \mathsf{Ling}(\mathcal{E}, \Phi) \mid \mathcal{L}_1 \vee \mathcal{L}_2$$

$$\text{Linguistic pattern } \mathcal{E} := \mathcal{E}^+ \mid \mathcal{E}^* \mid \mathcal{E}? \mid \mathcal{E}_1\ \mathcal{E}_2 \mid \ell \colon \mathcal{E} \mid \mathsf{Word} \mid \mathsf{ConstWord}(s) \mid \mathsf{ConstPhrase}(s_1, \ldots, s_k) \mid \mathsf{Syn}(s) \mid \mathsf{POS}(p) \mid \mathsf{Entity}(e) \mid \mathsf{NP} \mid \ldots$$

$$\text{Linguistic constraint } \Phi := \mathsf{SameSentence}(\ell_1, \ell_2) \mid \mathsf{Regex}(\ell, s) \mid \ldots \mid \Phi_1 \wedge \Phi_2 \mid \Phi_1 \vee \Phi_2 \mid \neg\Phi \mid \mathsf{true} \mid \mathsf{false}$$

$$\text{String } s := w \mid v_k \qquad \text{Part of speech } p \in \{\mathsf{Noun}, \mathsf{Verb}, \mathsf{Prep} \ldots\}$$

$$\text{ID labels } \ell, \eta := w \qquad \text{Entity type } e \in \{\mathsf{Person}, \mathsf{Org}, \mathsf{Place} \ldots\}$$

Visual patterns: webpage layout, colors, style, HTML, CSS
Describe stylistic webpage properties, *as seen by end-users*
Interpretation: rendering & DOM analysis

Structural patterns: implicit content schema, tables, lists
Describe relational patterns on *implicit tables*
Interpretation: table detection, plaintext analysis using PBE [1]

Linguistic patterns: text syntax, semantics, language, regexes
Describe fuzzy semantic subexpressions of the webpage text.
Interpretation: POS tagging, sentence parsing, entity recognition [2-5], synonymy detection [6]

LaSEWeb Search Algorithm

Given:

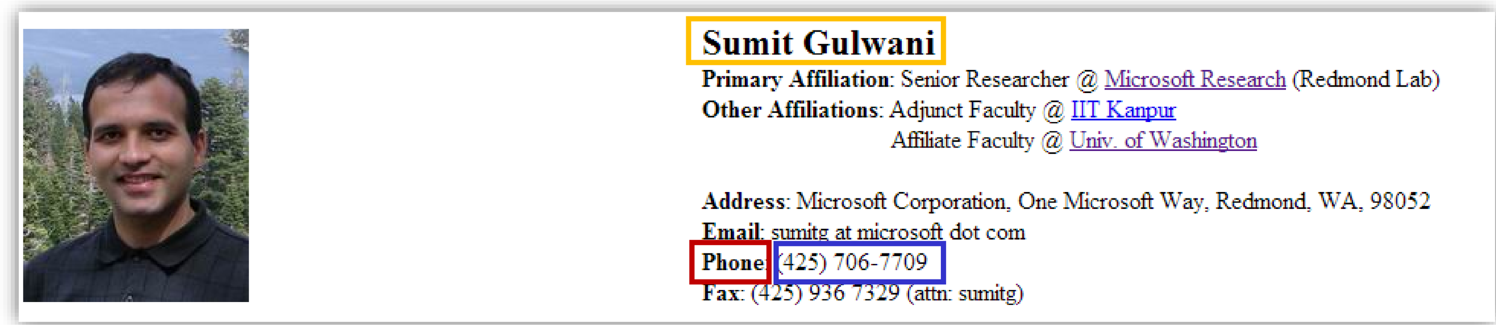
- Seed query function $q(\vec{v})$ — "email" ↦ "email inventor", …
- Similarity metric $\sigma(s_1, s_2)$ — "Sumit Gulwani" ≈ "Gulwani, S."
- LaSEWeb query $\mathcal{Q}$ — see example ⟶
- Answer label ℓ_a — a subexpression of $\mathcal{Q}$ to extract

Do:

1. Search the Web for top-k relevant webpages using $q(\vec{v})$.
2. Match the LaSEWeb query on each webpage and extract ℓ_a.
3. Cluster the resulting answer candidates based on similarity σ.
4. Rank the clusters and select representative answers.

```
function SEARCH(P = ⟨q, σ, Q, ℓ_a⟩, v⃗)
    U ← the results of Bing on q(v⃗)
    Substitute v_k in Q with values from v⃗
    C ← ∅              // set of clusters, C_i = {⟨s_k, {u_j}_{j=1}^{n_ik}⟩}_{k=1}^{m_i}
    for all URLs u_j ∈ U do
        N ← the <body> node of u_j
        S_j ← {M[ℓ_a] | M is the result of executing Q on N}
        for all answer strings s_k ∈ S_j do
            C_j ← {⟨s_k, {u_j}⟩}
            for all C ∈ C such that ∃s′ ∈ C: σ(s_k, s′) do
                Merge C_j with C
            C ← C ∪ {C_j}
    for all final clusters C_i ∈ C do
        a_i ← the most frequent string representation s_k ∈ C_i
        β_i ← (1/|U|) Σ_{j=1}^{|U|} Σ_{s∈C_i} c(s,u_j)/|S_j|    where:
              c(s, u_j) = # of times s was found at URL u_j
        U_i ← union of all source URLs for all s_k ∈ C_i
        yield return ⟨a_i, β_i, U_i⟩
```

Example



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

$$\mathcal{Q} = FW(Union(\eta_t\colon Leaf(v_1), \eta_b\colon S_b), \Psi)$$
$$\Psi = Layout(\eta_t, \eta_b, \mathrm{Down}) \wedge Nearby(\eta_t, \eta_b) \wedge Emphasized(\eta_t)$$
$$S_b = AttributeLookup(Syn(\text{"phone"}), \mathcal{L}_a)$$
$$\mathcal{L}_a = Ling(\ell_a, Regex(\ell_a, \texttt{"\textbackslash(\textbackslash d+\textbackslash)\textbackslash W*\textbackslash d+\textbackslash W*\textbackslash d+"}))$$

Evaluation

Micro-segments: 100,000+ user queries across 7 micro-segments from Bing search logs. Precision evaluated through random sampling, 95% in top-3 results. Average execution time: 5 sec/page.
Batch data extraction: 5 academic Web mining scenarios, precision and recall evaluated manually.

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%**

* The first author did this work during an internship at Microsoft Research Redmond.

[1] S. Gulwani. Automating string processing in spreadsheets using input-output examples. In POPL, 2011.
[2] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In ACL, 2005.
[3] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In ACL, 2003.
[4] 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.
[5] 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.
[6] W.-t. Yih, G. Zweig, and J. C. Platt. Polarity inducing latent semantic analysis. In ACL, 2012.