Automatic extraction of synonyms in a dictionary

Vincent D. Blondel Pierre P. Senellart

Université catholique de Louvain

UCL

1794 1994

Ecole Normale Supérieure

April 13th 2002

The dictionary graph

Computation (n.) The act or process of computing; calculation; reckoning.

Computation (n.) The result of computation; the amount computed.

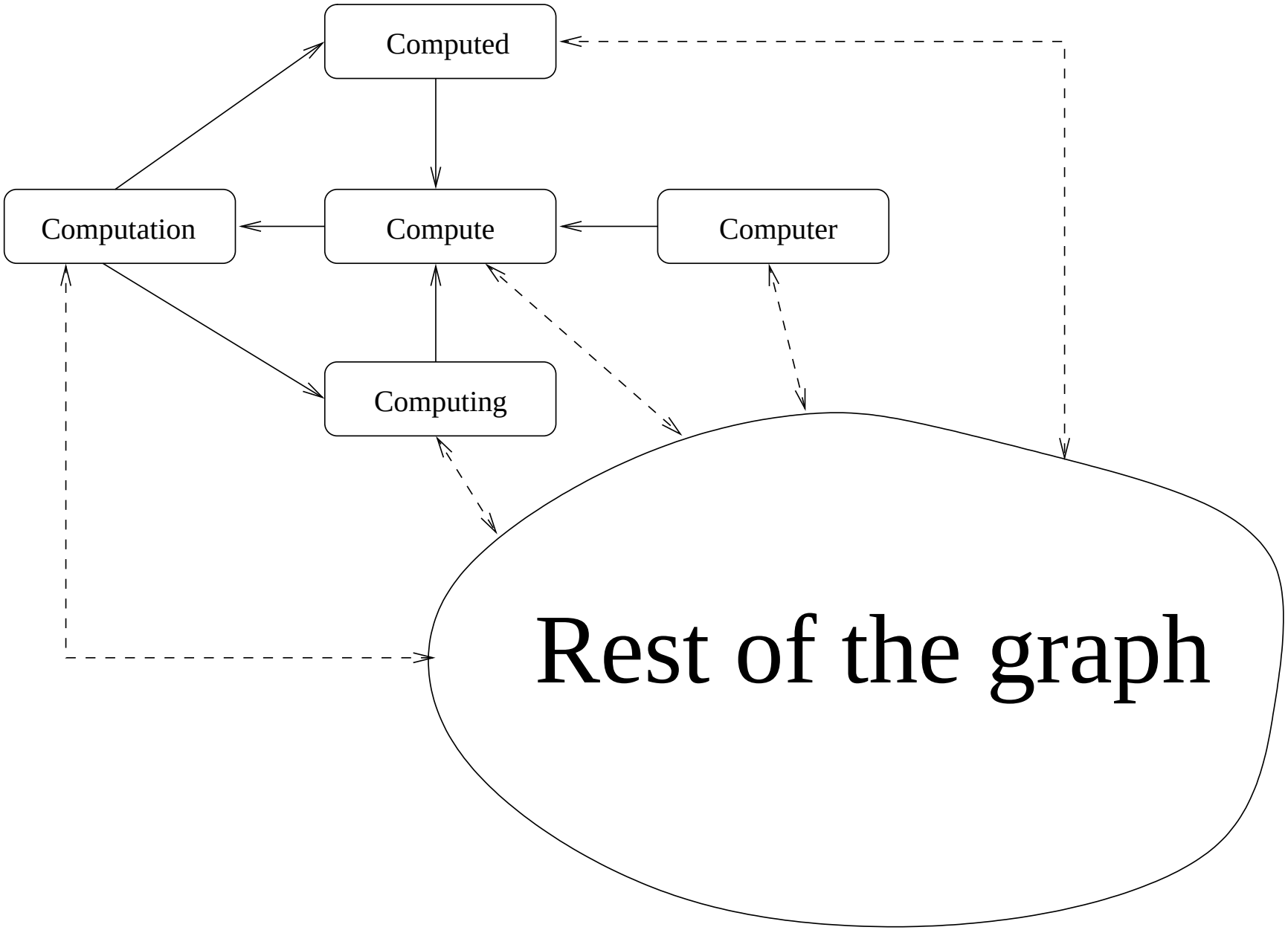
Computed (imp. & p. p.) of Compute

Computing (p. pr. & vb. n.) of Compute

Compute (v.t.) To determine calculation; to reckon; to count.

Compute (n.) Computation.

Computer (n.) One who computes.


Computed
Computation
Compute
Computer
Computing
Rest of the graph

Looking for near-synonyms

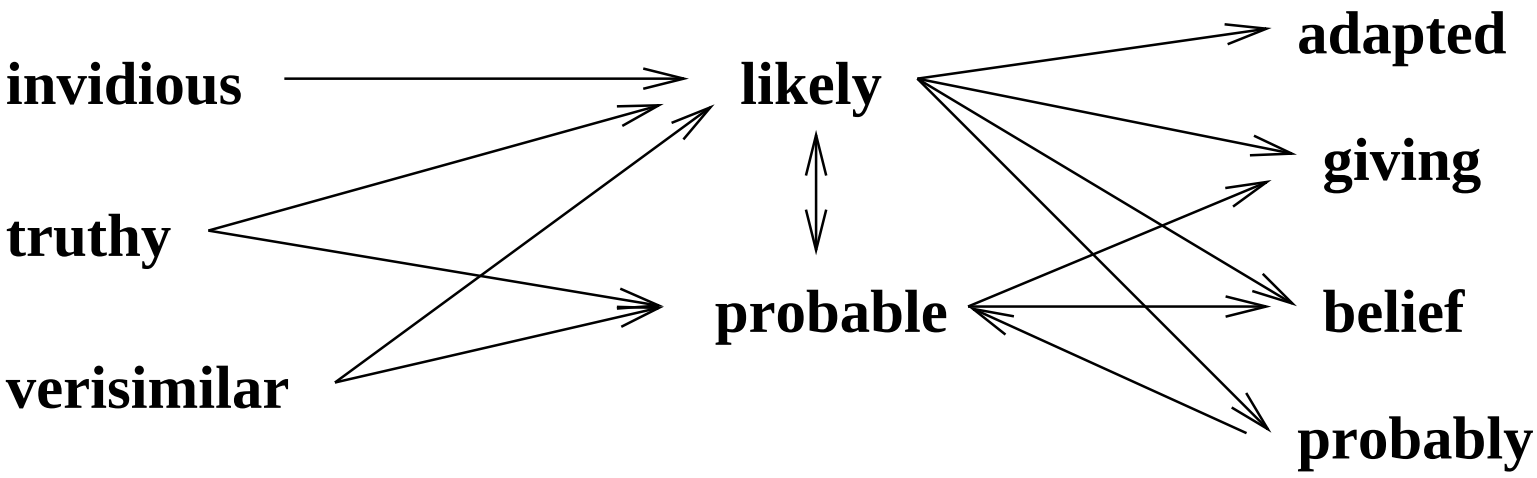
Definition. *The* neighborhood graph *of a node i in a directed graph G is the subgraph consisting of i, all parents of i and all children of i.*

i is some word we want a synonym of.

A will be the adjacency matrix of the neighborhood graph of i in the dictionary graph.

n is the order of A.

Subgraph of the neighborhood graph of likely



Hubs and Authorities on the Web

The Web as a graph:

- vertices=web pages
- edges=links

$$Hub \longrightarrow Authority$$

A *mutually reinforcing relationship*: good hubs are pages that point to good authorities and good authorities are pages pointed by good hubs.

Kleinberg's algorithm

x_i^1: iteratively computed *hub weights*
x_i^2: iteratively computed *authority weights*

$$x_0^1 = x_0^2 = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} x^1 \\ x^2 \end{pmatrix}_{t+1} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}_t$$

$$t = 0, 1, \dots$$

The principal eigenvectors of $A^T A$ and AA^T give respectively the *authority weights* and *hub weights* of the vertices of the graph.

An extension of Kleinberg's algorithm

Let $M(m,m)$ and $N(n,n)$ be the transition matrices of two oriented graphs.

Let $C = M \otimes N + M^T \otimes N^T$ where $\otimes$ is the Kronecker tensorial product.

We assume that the greatest eigenvalue of C is strictly greater than the absolute value of all other eigenvalues.

Then, the normalized principal eigenvector X of C gives the "similarity" between a vertex of M and a vertex of N: $X_{i \times n+j}$ characterizes the similarity between vertex i of M and vertex j of N.

In particular, if $M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, the result is that of Kleinberg's algorithm.

Application to the search for synonyms

$$1 \longrightarrow 2 \longrightarrow 3$$

We are looking for vertices "like 2" in the neighborhood graph of i.

Let $C = M \otimes A + M^T \otimes A^T$ where $M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$.

The principal eigenvector of C gives the similarity between a node in G and a node in the graph $1 \longrightarrow 2 \longrightarrow 3$.

We just select the subvector corresponding to the vertex 2 in order to have synonymy weights.

The vectors method

For each $1 \leq j \leq n, j \neq i$, compute:

$$\|(A_{i,.} - A_{j,.})\| + \|(A_{.,i} - A_{.,j})^T\|$$

(where $\|\ \|$ is some vector norm, $A_{i,.}$ and $A_{.,i}$ are respectively the ith line and the ith column of A).

For instance, if we choose the Euclidean norm, we compute:

$$\left(\sum_{k=1}^{n}(A_{i,k} - A_{j,k})^2\right)^{\frac{1}{2}} + \left(\sum_{k=1}^{n}(A_{k,i} - A_{k,j})^2\right)^{\frac{1}{2}}$$

The lower this value is, the better j is a synonym of i.

ArcRank

PageRank (Google): stationary distribution of weights over vertices corresponding to the principal eigenvector of the adjacency matrix.

ArcRank:

$$r_{s,t} = \frac{p_s/|a_s|}{p_t}$$

$|a_s|$ is the outdegree of s.

p_t is the pagerank of t.

The best synonyms of i are the other extremity of the best-ranked arcs arriving to or leaving from i.

Extraction of the graph

- Multiwords (e.g. All Saints', Surinam toad)
- Prefixes and suffixes (e.g un-, -ous)
- Different meanings of a word
- Derived forms (e.g. daisies, sought)
- Accentuated characters (e.g. proven/al, cr/che)
- Misspelled words

$112,169$ vertices - $1,398,424$ arcs.

Lexical units

$13,396$ lexical units not defined in the dictionary:

- Numbers (e.g. 14159265, 14th)
- Mathematical and chemical symbols (e.g. x3, fe3o4)
- Proper nouns (e.g. California, Aaron)
- Misspelled words (e.g. aligator, abudance)
- Undefined words (e.g. snakelike, unwound)
- Abbreviations (e.g. adj, etc)

Too frequent words

of	68187
a	47500
the	43760
or	41496
to	31957
in	23999
as	22529
and	16781
an	14027
by	12468
one	12216
with	10944
which	10446

Parts of speech

305 different categories transformed into combinations of:

- noun
- adjective
- adverb
- verb
- other (articles, conjunctions, interjections. . .)

Disappear

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	**vanish**	**vanish**	**epidemic**	**vanish**	**vanish**
2	**wear**	**pass**	**disappearing**	**go away**	**cease to exist**
3	**die**	**die**	**port**	**end**	**fade away**
4	**sail**	**wear**	**dissipate**	**finish**	**die out**
5	**faint**	**faint**	**cease**	**terminate**	**go**
6	**light**	**fade**	**eat**	**cease**	**evaporate**
7	**port**	**sail**	**gradually**		**wane**
8	**absorb**	**light**	**instrumental**		**expire**
9	**appear**	**dissipate**	**darkness**		**withdraw**
10	**cease**	**cease**	**efface**		**pass away**
Mark	3.6	6.3	1.2	7.5	8.6
Std dev.	1.8	1.7	1.2	1.4	1.3

Table 1: Near-synonyms for **disappear**

Parallelogram

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	**square**	**square**	**quadrilateral**	**quadrilateral**	**diamond**
2	**parallel**	**rhomb**	**gnomon**	**quadrangle**	**lozenge**
3	**rhomb**	**parallel**	**right-lined**	**tetragon**	**rhomb**
4	**prism**	**figure**	**rectangle**		
5	**figure**	**prism**	**consequently**		
6	**equal**	**equal**	**parallelopiped**		
7	**quadrilateral**	**opposite**	**parallel**		
8	**opposite**	**angles**	**cylinder**		
9	**altitude**	**quadrilateral**	**popular**		
10	**parallelopiped**	**rectangle**	**prism**		
Mark	4.6	4.8	3.3	6.3	5.3
Std dev.	2.7	2.5	2.2	2.5	2.6

Table 2: Near-synonyms for **parallelogram**

Sugar

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	**juice**	**cane**	**granulation**	**sweetening**	**darling**
2	**starch**	**starch**	**shrub**	**sweetener**	**baby**
2	**cane**	**sucrose**	**sucrose**	**carbohydrate**	**honey**
4	**milk**	**milk**	**preserve**	**saccharide**	**dear**
5	**molasses**	**sweet**	**honeyed**	**organic compound**	**love**
6	**sucrose**	**dextrose**	**property**	**saccarify**	**dearest**
7	**wax**	**molasses**	**sorghum**	**sweeten**	**beloved**
8	**root**	**juice**	**grocer**	**dulcify**	**precious**
9	**crystalline**	**glucose**	**acetate**	**edulcorate**	**pet**
10	**confection**	**lactose**	**saccharine**	**dulcorate**	**babe**
Mark	3.9	6.3	4.3	6.2	4.7
Std dev.	2.0	2.4	2.3	2.9	2.7

Table 3: Near-synonyms for **sugar**

Science

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	**art**	**art**	**formulate**	**knowledge domain**	**discipline**
2	**branch**	**branch**	**arithmetic**	**knowledge base**	**knowledge**
3	**nature**	**law**	**systematize**	**discipline**	**skill**
4	**law**	**study**	**scientific**	**subject**	**art**
5	**knowledge**	**practice**	**knowledge**	**subject area**	
6	**principle**	**natural**	**geometry**	**subject field**	
7	**life**	**knowledge**	**philosophical**	**field**	
8	**natural**	**learning**	**learning**	**field of study**	
9	**electricity**	**theory**	**expertness**	**ability**	
10	**biology**	**principle**	**mathematics**	**power**	
Mark	3.6	4.4	3.2	7.1	6.5
Std dev.	2.0	2.5	2.9	2.6	2.4

Table 4: Near-synonyms for **science**

Perspectives

- Extension of the subgraph
- Other dictionaries, other languages
- Other applications of the extension of Kleinberg's algorithm