# Cross-lingual and temporal Wikipedia analysis

### Göbölös-Szabó Julianna

MTA SZTAKI Data Mining and Search Group

June 14, 2013

Supported by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956)

伺 ト イヨト イヨ

通 とう ほうとう ほうど

### 1 Link prediction on multilingual Wikipedia

- Motivation About SimRank Simrank for multilingual Wikipedia Link prediction
- 2 Temporal Wikipedia search by edits and linkage Motivation Selecting temporal changing subgraph Personalized PageRank and Personalized HITS

# Section 1

## Link prediction on multilingual Wikipedia

æ

# Multilingual Wikipedia

イロト イヨト イヨト イヨト

æ



	Notice deroos				
icell Mahahana	La mandiana el Dellas nel un scolorge hamanistica internazione d'ana Della.				
An Apparture No. Article	Sense yours				
Collina Participa da Collinatado Registra Registra Registra Registra	Thermannee can be under a first the Assessment of the second				
bib a set					
Alto tepes en Cast fong Date Date Date Date Date Date Date	Ordering on service Understanding of the end of the en				
holds Magar	Origina press				

	Erdős-Zahl					
suptselte ber Wikipedia semerportale an A bis Z dälliger Artikel	De close 220 I secrete de Trainer o Dephen de l'assemblishe bacego e d'a Matematike Par Erabi, no copera evente de patientes evendente Autoria en la Colora imposieren, a cuatoria bano para de la Matematike Par Erabi, a como come a Matematike Defenso en Este 221 per a el vencomentaria en la de Augusta esta esta esta esta esta esta esta e	8	88			
Mitmachen Neuen Artikel anlegan Autorenportal Hilfe	Genika der Orbitrisine der Saltis-Zahl van Brauf Statis seitst dar Saltis-Zahl 3. dar Konsteiner, mit weitstern er problem Ha, hande die Ersbis-Zahl 1. Autoren, der Im Koussiever von Professiog bereichen im Ersbis saltist grauterin Hann, hande die Salt-Zahl 2. usu. Wonn han Werbranzg in isset fam zur einer Person henstelber zic, ist seme Ersbis-Zahl unerdiche, Es zagi schl, dass die Ersbis-Zahl der mötelsen Persone ernebet reverlich notes reter Hann in – de Bild. 2009 Personer mit einer wolfcher Wert under einer darschaftlichen Ersbis-Zahl der mötelber der Saltist van die mit auf der Bild. 2004 Ersbis-Zahl unerdiche, Es zagi schl, dass die Ersbis-Zahl der mötelber Vertragen in sone Ersbis vier einer der Saltist Wert in einer darschaftlich keiner der aufschnichtlichen Ersbis-Zahl der mötelber vertragen zur einer Bilder bestehen Als 2009 Personer mit einer wolfstehe Wert bestehen einer darschaftlichtlichen Ersbis-Zahl der mötelber bestehen ander einer wolfstehe Wert bestehen einer darschaftlichtlichen Ersbis-Zahl der mötelber bestehen ander darschaftlichtlichtlichtlichtlichtlichtlichtlich	Alce losteorieri mit Exdős und erhált ő- so die Erdős-Zahl 1. Wenn Bob nie mit Exdős, öber mit Alce kolaborieri hat, dann erhált er die Erdős-Zahl 2.				
Letzte Änderungen Kontakt	der Mathematik gearbeitet und mit über 500 verschiedenen Wissenschaftlern gemeinsam publiziert hat. Als Grundlage für die Berechnung der jeweiligen Endős- Zahl denen bibliografische Datenbanken, die vom Endős-Zahl-Projekt verwaltet und regelmäßig aktualisiert werden. Auf deren Basis lässt sich die Endős- Zahl gemeinter bibliografische Datenbanken, die vom Endős-Zahl-Projekt verwaltet und regelmäßig aktualisiert werden. Auf deren Basis lässt sich die Endős-	utors best	immen.			
Spenden	Von Bodeutung hir den einzelnen Mathematiker ist die Berechnung der Entdis-Zahl richt, auch wenn der Greph der Koastoerschaft zu Endis oft als Beispel für Graphen von Netzwerken in wissenschaftlichen Publikationen beruzzt wist. Abgerein berachtet veranschaulicht die Endis-Zahl einen Aspekt sozialer Netzwerke, der auch im Rahmen des Kreine-Weit-Phänomens behandeit wi					
Drucken/exportieren						
Verkzeuge	Andere verwandschaltsbezentrigen assen sich anlagig demenen, promientisses antiches besprei ist die Sacon-zah, die deer Konadonationen unter Schaltspreier	in deimer	151.			
anderen Sprachen	Einzelnachweise (Bearbeiter)					
ner Carada Česky	1. · Geffmen, Casper: And what is your Endos number? @ In: American Mathematical Monthly: 76, 1969. 2. · Facts about Eddls Annahoes and the Colaboration Graph. @ The distribution of Endos numbers. In: The Endos Number Project. Oakland University, 15. Septem 3. Ontaker ball Original, Determines Jul 2004.	nber 2010	. abgerufen	am		
EXAmples de	Weblinks (bearbeirer)					
Inglish	Webnitsenz des Estifs. Zahl-Dinielises 42 (proviliseth)					
uomi						
	Berechnung von Endős-Zahlen bei Mathscinet g					

Wikipedia articles about Erdős-number in German, French and Hungarian

# Multilingual Graph model

向下 イヨト イヨト

æ



Edge types:

- links between articles
- category-contains-article relationship
- category-hierarchy-links
- interwiki links (between languages)



- 3 languages: German, French, Hungarian
- snapshot from March 2012

lang.	articles	categories		
De	2 338 795	139 844		
Fr	2 408 097	199 708		
Hu	339 041	34 653		

Parallel articles

Parallel categories

De-Fr	482 196	De-Fr	22 175
De-Hu	108 949	De-Hu	4 840
Fr-Hu	119 559	Fr-Hu	5 387

- Only a small fraction of pages has an equivalent version
- Category hierarchies are entirely different

・ 同 ト ・ ヨ ト ・ ヨ ト

### Motivation:

- cleansing, expanding local Wikipedia:
  - new content from a bigger Wikipedia to a smaller
  - more detailed content from a smaller, better specified Wikipedia to the bigger one
- Tag recommendation in similarly structured networks (LibraryThing, Amazon)

We were focusing on:

- interwiki link recommendation for categories
- category recommendation for articles
- related entity recommendation for articles

Similar methods are used:

- 1 Setting candidates
- 2 Ranking candidates (with Jaccard, SimRank, etc.)

# **Basic SimRank Equation**

向下 イヨト イヨト

- "Two pages are similar if pointed to by similar pages" (Jeh-Widom KDD 2002)
- The similarity between objects a and b: sim(a, b) ∈ [0, 1]



$$sim(a, b) = \begin{cases} 1 & \text{if } a = b \\ \frac{C}{|N(a)| \cdot |N(b)|} \sum_{i=1}^{|N(a)|} \sum_{j=1}^{|N(b)|} sim(N_i(a), N_j(b)) & \text{otherwise} \end{cases}$$

- Similarity between *a* and *b* is the **average similarity** between in-neighbors of *a* and in-neighbors of *b*
- C is called **decay factor**, it is a constant between 0 and 1

向下 イヨト イヨト

**Expected meeting distance** is the expected time of how soon two random surfers (starting from a and from b) meet at the same node, walking **backwards** on edges.

$$EMD(a, b) = \sum_{v,l} P(after \ l \ steps \ a \ and \ b \ meet \ at \ v) \cdot l$$

Expected *f*-meeting distance

$$f - EMD(a, b) = \sum_{v,l} P(after \ l \ steps \ a \ and \ b \ meet \ at \ v) \cdot f(l)$$

Usually  $f(x) = C^x$  is choosen with  $C \in (0, 1)$ , since it transformes distance to similarity.

向下 イヨト イヨト

### Let's define

$$s(a,b) = \sum_{v,l} P( ext{after } l ext{ steps } a ext{ and } b ext{ meet at } v) \cdot C'$$

• It is easy to show that sim(a, b) is the same as s(a, b)

**Corollary:** SimRank can be approximated with (backwards) random walks.

### Random walk:

- 1. decide, whether we continue the walk
  - on a "normal" edge (with lpha probability)
  - or on an interwiki link (with  $1 \alpha$  probability).
- 2. select uniformly an edge with the type determined above

### Equivalent:

generating random walk on an edge-weighted graph

向下 イヨト イヨト

# SimRank for edge-weighted graphs



• 3 >

A⊒ ▶ ∢ ∃

Э

### SimRank for edge-weighted graphs



We choose according to the following probabilities. Let's go to D!

### SimRank for edge-weighted graphs



Standing in *D* we have the following oportunities.

문 문 문

Given German and French Wikipedia and we want to find a new category for article  $A_2$ 



3 1 3

Given German and French Wikipedia and we want to find a new category for article  $A_2$ 



**1** Take  $B_1$ , the equivalent article in German

Given German and French Wikipedia and we want to find a new category for article  $A_2$ 



- **1** Take  $B_1$ , the equivalent article in German
- 2 Take the categories of B<sub>1</sub> but discard trivial ones (K<sub>1</sub>'s equivalent is already the category of A<sub>2</sub>, K<sub>4</sub> doesn't have a pair in French)

Given German and French Wikipedia and we want to find a new category for article  $A_2$ 



- **1** Take  $B_1$ , the equivalent article in German
- 2 Take the categories of B<sub>1</sub> but discard trivial ones (K<sub>1</sub>'s equivalent is already the category of A<sub>2</sub>, K<sub>4</sub> doesn't have a pair in French)
- **3** The candidates are their French equivalents:  $C_1, C_3$

Given German and French Wikipedia and we want to find a new category for article  $A_2$ 



- **1** Take  $B_1$ , the equivalent article in German
- 2 Take the categories of B<sub>1</sub> but discard trivial ones (K<sub>1</sub>'s equivalent is already the category of A<sub>2</sub>, K<sub>4</sub> doesn't have a pair in French)
- **3** The candidates are their French equivalents:  $C_1, C_3$
- 4 Rank candidates

向下 イヨト イヨト

3

- Weighted Jaccard (details were skipped here)
- SimRank
- Novelty:

$$Nov(x) = 1 - SimRank(c_1, \ldots, c_n, x)$$

where x is a candidate category for article a, and the current categories of a are  $c_1, \ldots, c_n$ 

### Similarity of several nodes:

$$s(v_1,...,v_k) = \frac{C}{|I(v_1)|\cdots |I(v_k)|} \sum_{u_1 \in I(v_1)} \cdots \sum_{u_k \in I(v_k)} s(u_1,...,u_k)$$

# Evaluation

- In each experiment 10 % of the respective edges were deleted (Interwiki links: 13000, Categories: 1 914 000, related articles: 8.5 Mill.)
- For interwiki links: one ground truth for each input
- For categories and related articles: several ground truth instances
- Measures for the output quality:
  - MRR (mean reciprocial rank)
  - nDCG (standard measure for IR problems)
  - Recall
  - Precision
- Manual assessment for type-2 and type-3

This was a joint work with MPII, Saarbrücken (N. Prytkova, M.Spaniol, G.Weikum)

# Section 2

### Temporal Wikipedia search by edits and linkage

3

- Wikipedia has the great virtue of being utterly up-to-date
- A significant event usually has an immediate trace
- Considering a **chain of events**, we are often interested in the **causes and effects**, naturally represented by citations and links.
- If we want to know how a story evolved in time, we also need the information about the time of appearance of pages and links

向下 イヨト イヨト

## Change measure

		$\operatorname{Sep}$	Oct
		$\downarrow$	$\downarrow$
		Oct	Nov
	content	0.044	0.18
Muammar	inlink	0.55	0.12
Gaddafi	outlink	0.033	0.04
	total	0.63	0.34
Dooth of	content	0	7.71
Muommor	inlink	0	4.21
Coddofi	outlink	0	4.64
Gaudan	total	0	16.6
	content	7.78	0.79
Battle of	inlink	4.78	0.21
Sirte (2011)	outlink	4.9	0.14
	total	17.5	1.1
National	content	0.15	0.08
Transitional	inlink	0.91	0.13
Council	outlink	5.68	0.29
Council	total	6.7	0.5

We measure change as the sum of

- Difference between the logarithm of the **in-degree** between the two dates;
- Same for out-degree;
- Absolute difference between the number of words in the article between the two dates.

# Change measure

		Sep	Oct
		$\downarrow$	$\downarrow$
		Oct	Nov
	content	0.044	0.18
Muammar	inlink	0.55	0.12
Gaddafi	outlink	0.033	0.04
	total	0.63	0.34
Dooth of	content	0	7.71
Death of Museumon	inlink	0	4.21
Coddof	outlink	0	4.64
Gaddall	total	0	16.6
	content	7.78	0.79
Battle of	inlink	4.78	0.21
Sirte (2011)	outlink	4.9	0.14
	total	17.5	1.1
National	content	0.15	0.08
Transitional	inlink	0.91	0.13
Council	outlink	5.68	0.29
Council	total	6.7	0.5

We measure change as the sum of

- Difference between the logarithm of the **in-degree** between the two dates;
- Same for **out-degree**;
- Absolute difference between the number of words in the article between the two dates.
- The change of a node is interesting, if the neighborhood of the node has changed as well
- E.g. Learning to rank vs. Occupy movement

向下 イヨト イヨト

### Goal:

Given a query Q and we want to find a subgraph that consists of relevant articles respective to the topic, this graph changes with time and this change explains the considered events related to Q.

- 3 snapshots (2011. september, october, november)
- 35 queries with ground truth answers
- Evaluation:
  - recall, NDCG, MRR
  - graph density
- Visualizing the graphs with our in-house built tool (*subjective* evaluation)

### Main steps of our algorithm:

- 1 Composing the seed set from the search results
- 2 Expanding of seed set and consider the induced subgraph
- **3** Computing personalization vector
- **4** Assinging scores to the nodes
- Selecting the top-15 (and their induced subgraph in each snapshot)

向下 イヨト イヨト

- Wikipedia content is indexed by a search engine
- top results usually don't form a connected graph
- we expand this set of nodes in order to get a possibly connected component
- new nodes are expected to be relevant or recently edited
- candidates: 1-step neighborhood of the seed set

$$\mathsf{score}(v) = \max_{u \in \mathsf{seed}} \left( \mathit{IR}(u) + \frac{\mathit{change}(u) + \mathit{change}(v)}{2} \right)$$

expand with the nodes with highest rank

向下 イヨト イヨト

- both IR score and change are relevant:
- before combination both value need to be scaled
- $\alpha$ : trade-off between change and relevance
- *T*: depends on the distribution of change values (*T* = 10 worked fine)

$$p(u) = \alpha \cdot \frac{\mathsf{IR}(u)}{\mathsf{maxIR}} + (1 - \alpha) \cdot \frac{\mathsf{change}(u)}{(\mathsf{change}(u) + \mathcal{T})}$$

#### Random surfer model

- Browsing the web, following hyperlinks
- Sometimes she gets bored and teleports
- When teleporting, take distribution *d* (instead of uniform distribution)

$$PPR_d^{(k+1)T} = PPR_d^{(k)T}((1-\alpha)M + \alpha \cdot D)$$
$$= PPR_d^{(1)T}((1-\alpha)M + \alpha \cdot D)^k$$

where D has all rows equal to the personalization vector  $\mathbf{d} = (d_1, \dots, d_n)$ 

$$D = \begin{pmatrix} \mathbf{d} \\ \mathbf{d} \\ \dots \\ \mathbf{d} \end{pmatrix}$$

Idea behind HITS:

- A good hub is a page that pointed to many other pages,
- A good authority is a page that was linked by many different hubs

$$\hat{a}(v) = \sum_{vu \in E} w(vu) \cdot h(u), \quad a = \hat{a}/||a||_{\infty}$$
$$\hat{h}(v) = \sum_{uv \in E} w(uv) \cdot a(u), \quad h = \hat{h}/||h||_{\infty}$$

向下 イヨト イヨト

高 とう ヨン うまと

- **supersource:** a new node of the graph which is connected with each node of the original graph, and the weight of an edge corresponds to the importance of the respective node in the personalization
- The supersource distributes a fixed amount of score in each iteration

$$\hat{a}(v) = \sum_{vu \in E} w(vu) \cdot h(u) + c \cdot p(v), \quad a = \hat{a}/||a||_{\infty}$$
$$\hat{h}(v) = \sum_{uv \in E} w(uv) \cdot a(u) + c \cdot p(v), \quad h = \hat{h}/||h||_{\infty}$$

Multiple versions:

- Scores from Hubs
- Scores from Authorities
- combination of Hub and Authority vector

Other combinations are possible as well:

- Hubs & PageRank
- Authority & Pagerank
- Hubs & Authority & Pagerank

Results

Э



nDCG values

Growth of number of edges

・ロト ・回ト ・ヨト ・ヨト

# An example for changing subgraph



#### Result for "Greek economy"

# An example for changing subgraph



#### Result for "Greek economy"

(日) (四) (王) (王) (王)

# An example for changing subgraph



#### Result for "Greek economy"

(日) (四) (王) (王) (王)

# Thank you for your attention!

(4) (5) (4) (5) (4)

æ