PROJECT PERIODIC REPORT

Grant Agreement number: 288956

Project acronym: NADINE

Project title: New tools and Algorithms for Directed Network analysis

Funding Scheme: Small or medium-scale focused research project (STREP)

Periodic report: 1st X 2nd

Period covered: from 1.5.2012 to 31.10.2013

Name, title and organisation of the scientific representative of the project's coordinator[1]:

Dr. Dima Shepelyansky

Directeur de recherche au CNRS

Lab de Phys. Theorique, Universite Paul Sabatier, 31062 Toulouse, France

Tel: +331 5 61556068, Fax: +33 5 61556065, Secr.: +33 5 61557572

E-mail: dima@irsamc.ups-tlse.fr; URL: www.quantware.ups-tlse.fr/dima

Project website address: www.quantware.ups-tlse.fr/FETNADINE/

[1] Usually the contact person of the coordinator as specified in Art. 8.1. of the grant agreement

NADINE DELIVERABLE D4.1.

It is based on milestones M4, M7(in progress), M8(in progress). M13(in progress), with deliverable publications:

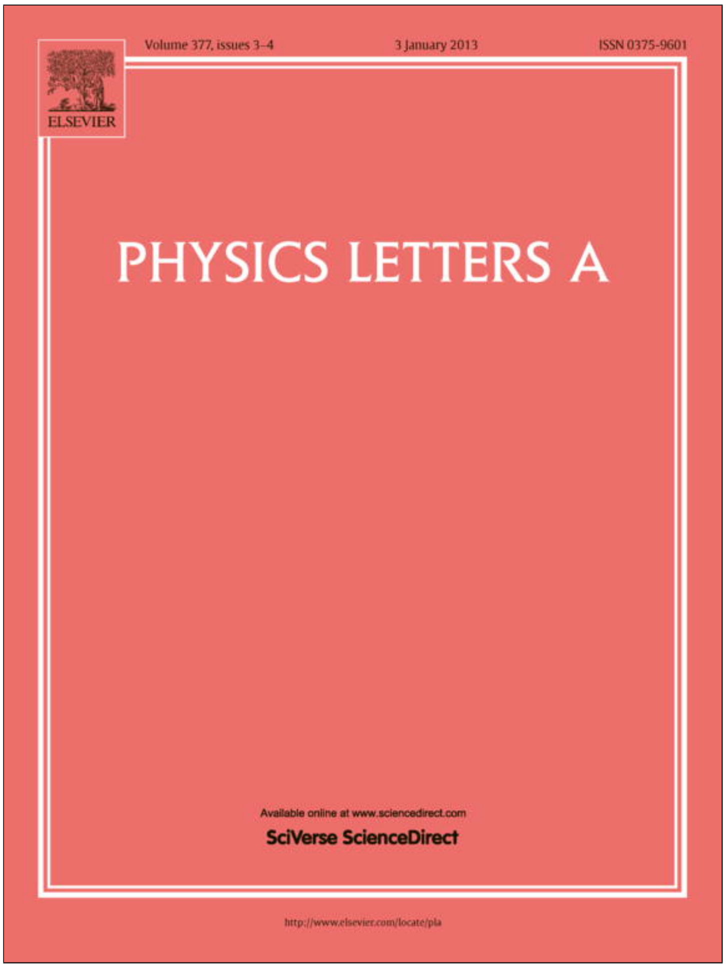
[2] P1.2 L.Ermann and D.L. Shepelyansky **"Ecological analysis of world trade"**, Phys. Lett. A v.377, p.250 (2013) (arXiv:1201.3584[q-fin.GN], 2012)

[10] P1.10 Y.-H.Eom and D.L. Shepelyansky, **"Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles"**, PLoS ONE v.8(10), p.e74554 (2013) (arXiv:1306.6259 [cs.SI], 2013)

[20] P3.2 A.Garzo, B.Daroczy, T.Kiss, D.Siklosi, and A.A.Benczur, **"Cross-Lingual Web Spam Classification"**, The 3rd Joint WICOW/AIRWeb Workshop on Web Quality in conj. WWW 2013, Rio de Janeiro, Brasil. May 13 (2013), Proceedigs of the 22nd international conference on World Wide Web companion

[21] P3.3 M.Erdelyi, A.A.Benczur, B.Daroczy, A.Garzo, T.Kiss and D.Siklosi, **"The classification power of Web features"**, Internet Mathematics, to appear (2013)

[23] P3.5 A.Garzo, A.A.Benczur, C.I.Sidlo, D.Tahara, E.F.Wyatt, **"Real-time streaming mobility analysis"**, Conference: IEEE Big Data 2013



Volume 377, issues 3–4 3 January 2013 ISSN 0375-9601

ELSEVIER

PHYSICS LETTERS A

Available online at www.sciencedirect.com

SciVerse ScienceDirect

http://www.elsevier.com/locate/pla

Physics Letters A 377 (2013) 250–256

Contents lists available at SciVerse ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Ecological analysis of world trade

L. Ermann [b,a], D.L. Shepelyansky [a,*]

[a] Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France
[b] Departamento de Física Teórica, GIyA, Comisión Nacional de Energía Atómica, Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 22 March 2012
Received in revised form 28 August 2012
Accepted 31 October 2012
Available online 21 November 2012
Communicated by A.R. Bishop

Keywords:
Complex networks
Nestedness
World trade

ABSTRACT

Ecological systems have a high complexity combined with stability and rich biodiversity. The analysis of their properties uses a concept of mutualistic networks and provides a detailed understanding of their features being linked to a high nestedness of these networks. Using the United Nations COMTRADE database we show that a similar ecological analysis gives a valuable description of the world trade: countries and trade products are analogous to plants and pollinators, and the whole trade network is characterized by a high nestedness typical for ecological networks. Our approach provides new mutualistic features of the world trade.

1. Introduction

Ecological systems are characterized by high complexity and biodiversity [1] linked to nonlinear dynamics and chaos emerging in the process of their evolution [2,3]. The interactions between species form a complex network whose properties can be analyzed by the modern methods of scale-free networks [4–7]. An important feature of ecological networks is that they are highly structured, being very different from randomly interacting species [7,8]. Recently it has been shown that the mutualistic networks between plants and their pollinators [8–12] are characterized by high nestedness [13–16] which minimizes competition and increases biodiversity. It is argued [14] that such type of networks appear in various social contexts such as garment industry [15] and banking [17,18]. Here we apply a nestedness analysis to the world trade network using the United Nations COMTRADE database [19] for the years 1962–2009. Our analysis shows that countries and trade products have relations similar to those of plants and pollinators and that the world trade network is characterized by a high nestedness typical of ecosystems [14]. This provides new mutualistic characteristics for the world trade.

2. Results

The mutualistic World Trade Network (WTN) is constructed on the basis of the UN COMTRADE database [19] from the matrix of trade transactions $M^{p}_{c',c}$ expressed in USD for a given product (commodity) p from country c to country c' in a given year (from 1962 to 2009). For product classification we use 3-digit Standard International Trade Classification (SITC) Rev. 1 with the number of products $N_p = 182$. All these products are described in [19] in the commodity code document SITC Rev. 1. The number of countries varies between $N_c = 164$ in 1962 and $N_c = 227$ in 2009. The import and export trade matrices are defined as $M^{(i)}_{p,c} = \sum_{c'=1}^{N_c} M^{p}_{c,c'}$ and $M^{(e)}_{p,c} = \sum_{c'=1}^{N_c} M^{p}_{c',c}$ respectively. We use the dimensionless matrix elements $m^{(i)} = M^{(i)}/M_{max}$ and $m^{(e)} = M^{(e)}/M_{max}$ where for a given year $M_{max} = max\{max[M^{(i)}_{p,c}], max[M^{(e)}_{p,c}]\}$. The distribution of matrix elements $m^{(i)}$, $m^{(e)}$ in the plane of indexes p and c, ordered by the total amount of import/export in a decreasing order, is shown in Fig. 1 for years 1968 and 2008 (years 1978, 1988, 1998 are shown in Fig. S-1 of Supporting Information (SI)). These figures show that globally the distributions of $m^{(i)}$, $m^{(e)}$ remain stable in time especially in a view of 100 times growth of the total trade volume during the period 1962–2009. The fluctuations of $m^{(e)}$ are visibly larger compared to $m^{(i)}$ case since certain products, e.g. petroleum, are exported by only a few countries while it is imported by almost all countries.

To use the methods of ecological analysis we construct the mutualistic network matrix for import $Q^{(i)}$ and export $Q^{(e)}$ whose matrix elements take binary value 1 or 0 if corresponding elements $m^{(i)}$ and $m^{(e)}$ are respectively larger or smaller than a certain trade threshold value μ. The fraction φ of nonzero matrix elements varies smoothly in the range $10^{-6} \leqslant \mu \leqslant 10^{-2}$ (see Fig. S-2 of SI) and the further analysis is not really sensitive to the actual μ value inside this broad range. Indeed, the variation of μ in

[*] Corresponding author.
E-mail address: ermann@tandar.cnea.gov.ar (L. Ermann).
URLs: http://www.tandar.cnea.gov.ar/~ermann (L. Ermann), http://www.quantware.ups-tlse.fr/dima (D.L. Shepelyansky).

0375-9601/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physleta.2012.10.056

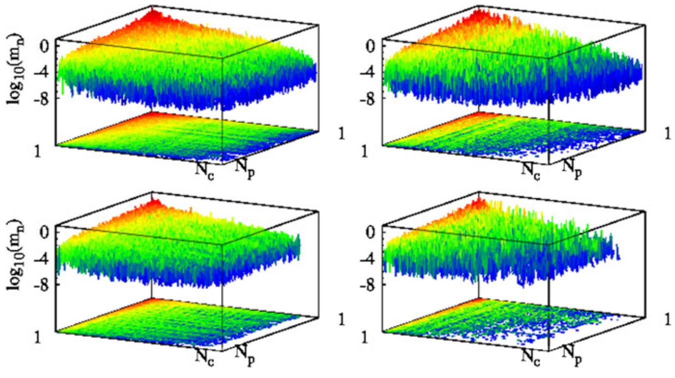


Fig. 1. Normalized import/export WTN matrix elements $m^{(i)}$ and $m^{(e)}$ shown on left/right panels for years 1968 (bottom) and 2008 (top). Each panel represents the dimensionless trade matrix elements $m^{(i)} = M^{(i)}/M_{max}$ and $m^{(e)} = M^{(e)}/M_{max}$ on a thee-dimensional (3D) plot as a function of indexes of countries and products. Here products/countries ($p = 1, \dots, N_p$ and $c = 1, \dots, N_c$) are ordered in a decreasing order of product/country total import or export in a given year. The color is proportional to the amplitude of the matrix element changing from red (for amplitude maximum) to blue (for zero amplitude). Each panel shows the 3D distribution and its projection on 2D plane of countries–products in which the amplitude of matrix elements is shown by the same color as in 3D. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

the range $10^{-5} \leqslant \mu \leqslant 10^{-3}$ by two orders of magnitude produces a rather restricted variation of φ only by a factor two.

It is important to note that in contrast to ecological systems [14] the world trade is described by a directed network and hence we characterize the system by two mutualistic matrices $Q^{(i)}$ and $Q^{(e)}$ corresponding to import and export. Using the standard nestedness BINMATNEST algorithm [20] we determine the nestedness parameter η of the WTN and the related nestedness temperature $T = 100(1 - \eta)$. The algorithm reorders lines and columns of a mutualistic matrix concentrating nonzero elements as much as possible in the top-left corner and thus providing information about the role of immigration and extinction in an ecological system. A high level of nestedness and ordering can be reached only for systems with low T. It is argued that the nested architecture of real mutualistic networks increases their biodiversity.

The nestedness matrices generated by the BINMATNEST algorithm [20] are shown in Fig. 2 for ecology networks ARR1 ($N_{pl} = 84$, $N_{anim} = 101$, $\varphi = 0.043$, $T = 2.4$) and WES ($N_{pl} = 207$, $N_{anim} = 110$, $\varphi = 0.049$, $T = 3.2$) from [12,21]. Using the same algorithm we generate the nestedness matrices of WTN using the mutualistic matrices for import $Q^{(i)}$ and export $Q^{(e)}$ for the WTN in years 1968 and 2008 using a fixed typical threshold $\mu = 10^{-3}$ (see Fig. 2; the distributions for other μ values have a similar form and are shown in Fig. S-3 of SI). As for ecological systems, for the WTN data we also obtain rather small nestedness temperature ($T \approx 6/8$ for import/export in 1968 and $T \approx 4/8$ in 2008 respectively). These values are by a factor 9/4 of times smaller than the corresponding T values for import/export from random generated networks with the corresponding values of φ.

The detailed data for T in all years are shown in Fig. 3 and the comparison with the data for random networks is given in Figs. S-4–S-6 in SI. The data of Fig. 3 show that the value of T changes by about 30–40% with variation of μ by a factor 1000. We think that this is relatively small variation of T compared to enormous variation of μ that confirms the stability and relevance of ecological analysis and nestedness ordering. The nestedness temperature T remains rather stable in time: in average there is 40% drop of T from 1962 to 2000 and 20% growth from 2000 to 2009. We attribute the growth in last decade to the globalization of trade. Even if the nestedness temperature T may be sensitive to

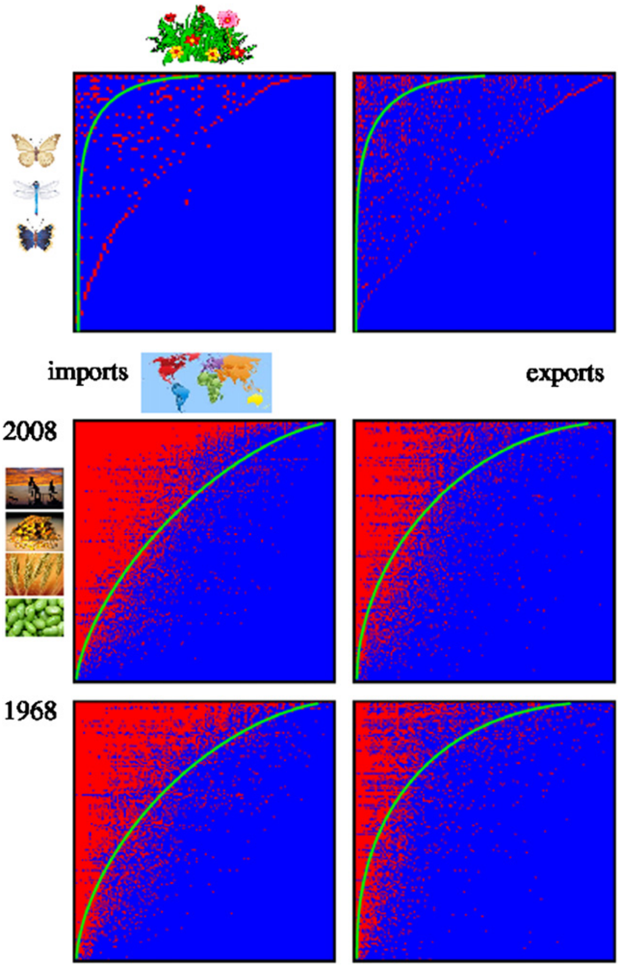


Fig. 2. Nestedness matrices for the plant–animal mutualistic networks on top panels, and for the WTN of countries–products on middle and bottom panels. Top-left and top-right panels represent data of ARR1 and WES networks from [12,21]. The WTN matrices are computed with the threshold $\mu = 10^{-3}$ and corresponding $\varphi \approx 0.2$ for years 1968 (bottom) and 2008 (middle) for import (left panels) and export (right panels). Red and blue represent unit and zero elements respectively; only lines and columns with nonzero elements are shown. The order of plants–animals, countries–products is given by the nestedness algorithm [20], the perfect nestedness is shown by green curves for the corresponding values of φ. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

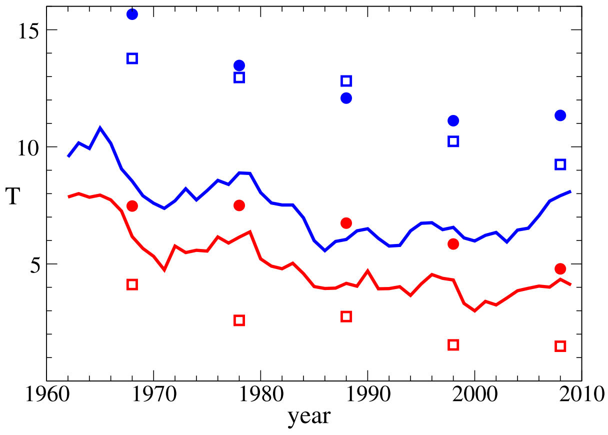


Fig. 3. Nestedness temperature T as a function of years for the WTN for $\mu = 10^{-3}$ (curves), 10^{-4} (circles), 10^{-6} (squares); import and export data are shown in red and blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

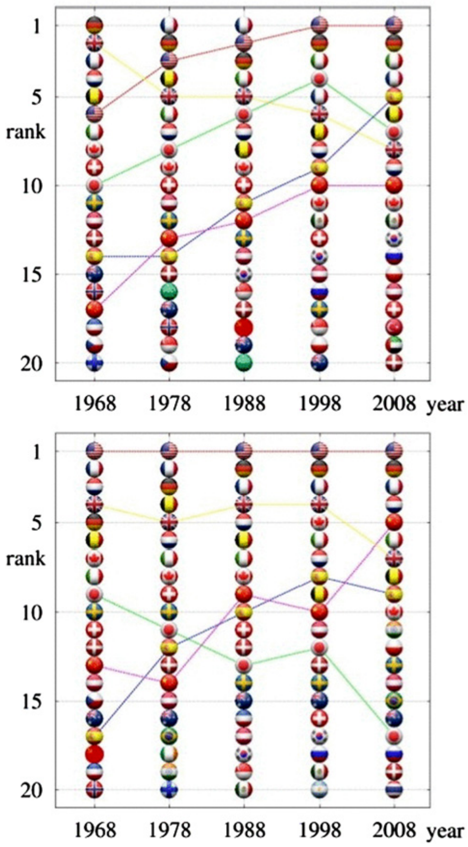


Fig. 4. Top 20 EcoloRank countries as a function of years for the WTN import/export on top/bottom panels. The ranking is given by the nestedness algorithm [20] for the trade threshold $\mu = 10^{-3}$; each country is represented by its corresponding flag. As an example, dashed lines show time evolution of the following countries: USA, UK, Japan, China, Spain. (For interpretation of the references to color in this figure, the reader is referred to the web version of this Letter.)

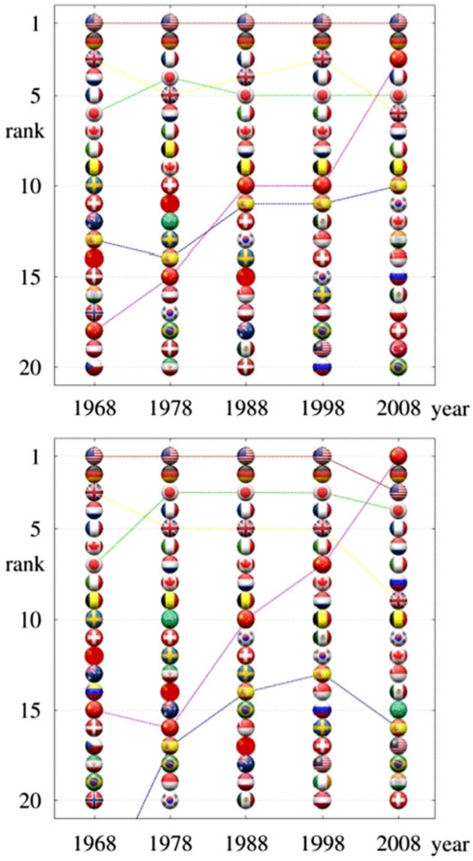


Fig. 5. Top 20 countries as a function of years ranked by the total monetary trade volume of the WTN in import/export on top/bottom panels respectively; each country is represented by its corresponding flag. Dashed lines show time evolution of the same countries as in Fig. 4.

variation of φ the data of Figs. S-2 and S-6 show that in the main range of $10^{-5} \leqslant \mu \leqslant 10^{-3}$ the variation of φ and T remains rather small. The comparison with the randomly generated networks also shows that they have significantly larger T values compared to the values found for the WTN (see also discussion of Figs. S-4–S-6 in SI).

The small value of nestedness temperature obtained for the WTN confirms the validity of the ecological analysis of WTN structure: trade products play the role of pollinators which produce exchange between world countries, which play the role of plants. Like in ecology the WTN evolves to the state with very low nestedness temperature that satisfies the ecological concept of system stability appearing as a result of high network nestedness [14].

The nestedness algorithm [20] creates effective ecological ranking (EcoloRanking) of all UN countries. The evolution of 20 top ranks throughout the years is shown in Fig. 4 for import and export. This ranking is quite different from the more commonly applied ranking of countries by their total import/export monetary trade volume [22] (see corresponding data in Fig. 5) or recently proposed democratic ranking of WTN based on the Google matrix analysis [23]. Indeed, in 2008 China is at the top rank for total export volume but it is only at 5th position in EcoloRanking (see Figs. 4, 5 and Table 1 in SI). In a similar way Japan moves down from 4th to 17th position while the USA raises up from 3rd to 1st rank.

The same nestedness algorithm generates not only the ranking of countries but also the ranking of trade products for import and export which is presented in Fig. 6. For comparison we also show there the standard ranking of products by their trade volume. In Fig. 6 the color of symbol marks the 1st SITC digit described in [19] and in Table 2 in SI.

3. Discussion

The origin of such a difference between EcoloRanking and trade volume ranking of countries is related to the main idea of mutualistic ranking in ecological systems: the nestedness ordering stresses the importance of mutualistic pollinators (products for WTN) which generate links and exchange between plants (countries for WTN). In this way generic products, which participate in the trade between many countries, become of primary importance even if their trade volume is not at the top lines of import or export. In fact such mutualistic products glue the skeleton of the world trade while the nestedness concept allows to rank them in order of their importance. The time evolution of this EcoloRanking of products of WTN is shown in Fig. 6 for import/export in comparison with the product ranking by the monetary trade volume (since the trade matrix is diagonal in product index the ranking of products in the latter case is the same for import/export). The top and middle panels have dominate colors corresponding to machinery (SITC 7; blue) and mineral fuels (3; black) with a moderate contribution of chemicals (5; yellow) and manufactured articles (8; cyan) and a small fraction of goods classified by material (6; green). Even if the global structure of product ranking by trade volume has certain similarities with import EcoloRanking there are also important new elements. Indeed, in 2008 the mutualistic significance of petroleum products (SITC 332), *machindus* (machines for special industries 718) and

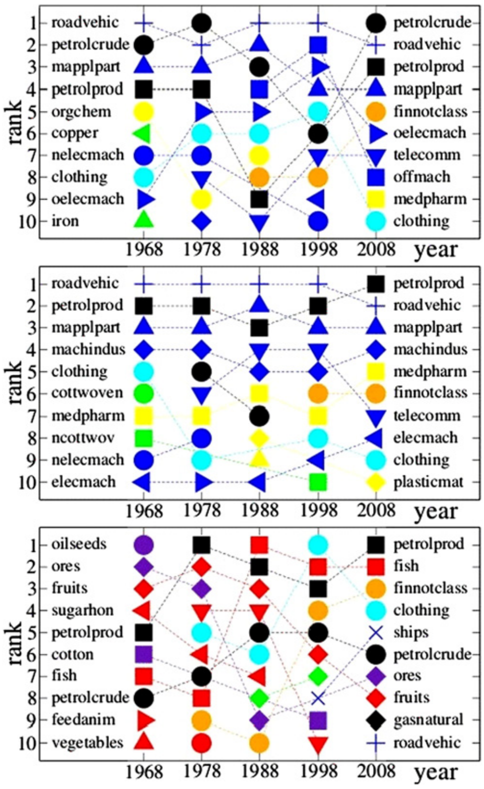


Fig. 6. Top 10 ranks of trade products as a function of years for the WTN. Top panel: ranking of products by monetary trade volume; middle/bottom panels: ranking is given by the nestedness algorithm [20] for import/export with the trade threshold $\mu = 10^{-3}$. Each product is shown by its own symbol with short name written for years 1968, 2008; symbol color marks 1st SITC digit; SITC codes of products and their names are given in Table 2 of SI. (For interpretation of the references to color in this figure, the reader is referred to the web version of this Letter.)

medpharm (medical–pharmaceutic products 541) is much higher compared to their volume ranking, while petroleum crude (331) and office machines (714) have smaller mutualistic significance compared to their volume ranking.

The new element of EcoloRanking is that it differentiates between import and export products while for trade volume they are ranked in the same way. Indeed, the dominant colors for export (Fig. 6, bottom panel) correspond to food (SITC 0; red) with contribution of black (present in import) and crude materials (2; violet), followed by cyan (present in import) and more pronounced presence of *finnotclass* (commodities/transactions not classified 9; brown). EcoloRanking of export shows a clear decrease tendency of dominance of SITC 0 and SITC 2 with time and increase of importance of SITC 3, 7. It is interesting to note that petroleum products SITC 332 is very vulnerable in volume ranking due to significant variations of petroleum prices but in EcoloRanking this product keeps the stable top positions in all years showing its mutualistic structural importance for the world trade. EcoloRanking of export shows also importance of fish (SITC 031), clothing (SITC 841) and fruits (SITC 051) which are placed on higher positions compared to their volume ranking. At the same time *roadvehic* (SITC 732), which are at top volume ranking, have relatively low ranking in export since only a few countries dominate the production of road vehicles.

It is interesting to note that in Fig. 6 petroleum crude is at the top of trade volume ranking e.g. in 2008 (top panel) but it is absent in import EcoloRanking (middle panel) and it is only on 6th position in export EcoloRanking (bottom panel). A similar feature is visible for years 1968, 1978. On a first glance this looks surprising but in fact for mutualistic EcoloRanking it is important that a given product is imported from top EcoloRank countries: this is definitely not the case for petroleum crude which practically is not produced inside top 10 import EcoloRank countries (the only exception is the USA, which however also does not export much). Due to that reason this product has low mutualistic significance.

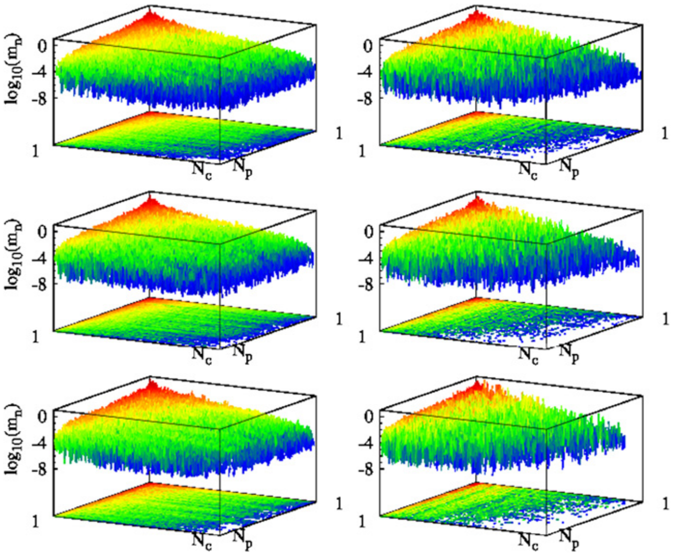


Fig. S-1. Same type of WTN matrix data as in Fig. 1 shown for years 1978, 1988, 1998 in panels from bottom to top respectively.

The mutualistic concept of product importance is at the origin of significant difference of EcoloRanking of countries compared to the usual trade volume ranking (see Figs. 4, 5). Indeed, in the latter case China and Japan are at the dominant positions but their trade is concentrated in specific products which mutualistic role is relatively low. In contrast the USA, Germany and France keep top three EcoloRank positions during almost 40 years clearly demonstrating their mutualistic power and importance for the world trade.

In conclusion, our results show the universal features of ecologic ranking of complex networks with promising future applications to trade, finance and other areas.

Acknowledgements

We thank Arlene Adriano and Matthias Reister (UN COMTRADE) for provided help and friendly access to the database [19]. This work is done in the frame of the EC FET Open project “New tools and algorithms for directed network analysis” (NADINE No. 288956).

Appendix A. Supporting information

Here we present the Supporting Information (SI) for the main part of the Letter, it includes Figs. S-1–S-6, Table 1, Table 2.

In Fig. S-1, in a complement to Fig. 1, we show the normalized WTN matrix for import $m^{(i)}$ and export $m^{(e)}$ at additional years 1978, 1988, 1998. As in Fig. 1 all products and countries are ordered in a decreasing order of product $(p = 1, \dots, N - p)$ and country $(c = 1, \dots, N_c)$ import (left panels) and export (right panels) in a given year. These data show that the global distribution remains stable in time: indeed, the global monetary trade volume was increased by a factor 100 from year 1962 to 2008 (see e.g. Fig. 5 in [20]) but the shape of the distribution remained essentially the same.

The dependence of the fraction φ of nonzero elements of the mutualistic matrices of import $Q^{(i)}$ and export $Q^{(e)}$ on the cutoff threshold μ is shown in Fig. S-2. In the range of $10^{-6} \leqslant \mu \leqslant 10^{-2}$ there is a smooth relatively weak variation of φ with μ.

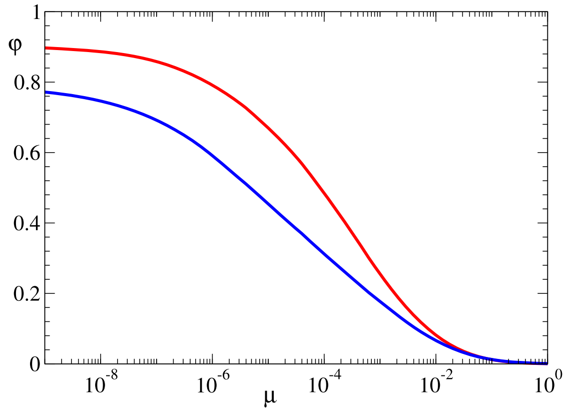


Fig. S-2. The fraction φ of nonzero matrix elements for the mutualistic network matrices of import $Q^{(i)}$ and export $Q^{(e)}$ as a function of the cutoff trade threshold μ for the normalized WTN matrices $m^{(i)}$ and $m^{(e)}$ for the year 2008; the red curve shows the case of import while the blue curve shows the case of export network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

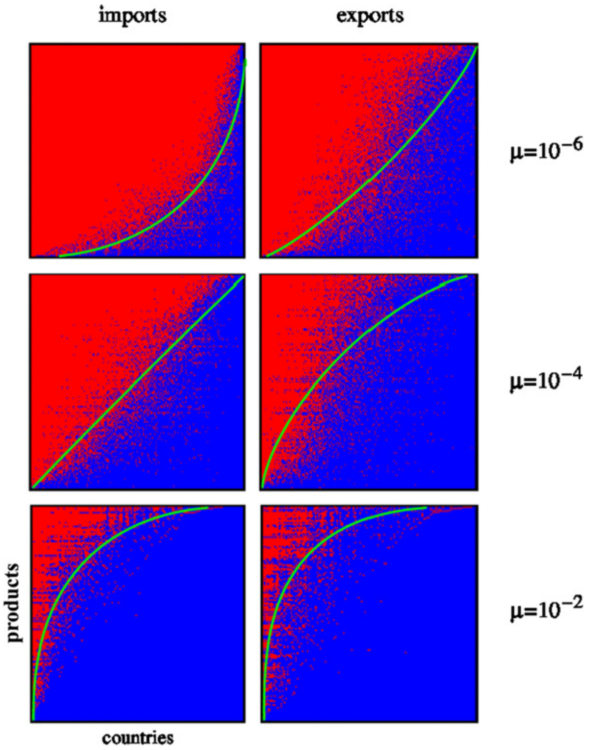


Fig. S-3. Same as in Fig. 2: nestedness matrix for the WTN data in 2008 shown for the threshold values $\mu = 10^{-6}, 10^{-4}, 10^{-2}$ (from top to bottom); the perfect nestedness is shown by green curves for the corresponding values of φ taken from Fig. S-2. (For interpretation of the reference to color in this figure legend, the reader is referred to the web version of this Letter.)

In Fig. S-3, in addition to Fig. 2, we show the nestedness matrices of WTN at various values of the cutoff threshold μ. The data at various μ values show that in all cases the nestedness algorithm [17] correctly generates a matrix with nestedness structure.

The variation of the nestedness temperature T with time is shown in Fig. 3 at several values of the trade threshold μ. These data show that in average the value of T for export is higher than for import. We attribute this to stronger fluctuations of matrix elements of $m^{(e)}$ compared to those of $m^{(i)}$ that is well visible in Figs. 1, S-1. As it is pointed in the main part, we attribute this to the fact that e.g. only a few countries export petroleum crude while the great majority of countries import this product.

In Fig. S-4 we show the nestedness temperature dependence on time for the case of random generated networks which have the same fraction of nonzero matrix elements φ as the WTN in the given year and $\mu = 10^{-3}$. These data, compared with those of Fig. 3, really demonstrate that the real WTN has values of T

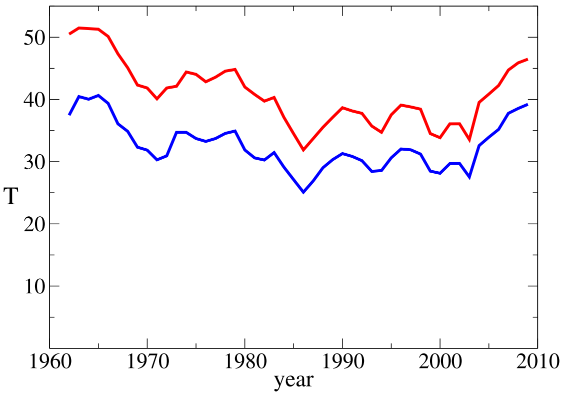


Fig. S-4. Nestedness temperature T for the model given by random generated networks; here T is computed with 500 random realizations of network for each year using N_p, N_c and φ of the corresponding WTN data in this year at $\mu = 10^{-3}$; import/export data are shown by red/blue curves respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

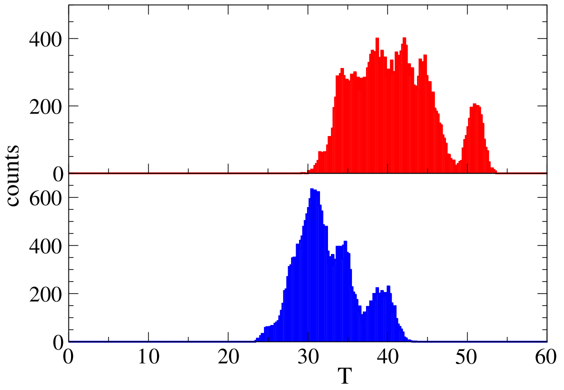


Fig. S-5. Histogram of temperatures for 500 random generated networks per year (from 1962 to 2009). Top (bottom) panel represents import (export) data; here the parameter values of N_p, N_c and φ are as for the corresponding WTN years at $\mu = 10^{-3}$.

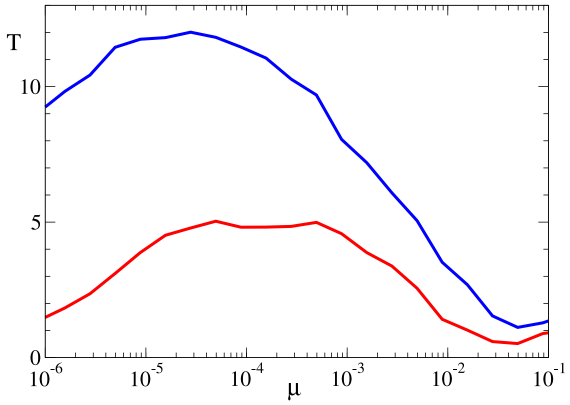


Fig. S-6. Nestedness temperature in the WTN for the year 2008 as a function of threshold μ; import/export networks are shown by red/blue curves respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

Table 1
Top 20 ranks of countries for import and export with ranking by the monetary trade volume and by the nestedness algorithm at two threshold values μ (year 2008).

Rank	Import			Export		
	Money	$\mu = 10^{-3}$	$\mu = 10^{-2}$	Money	$\mu = 10^{-3}$	$\mu = 10^{-2}$
1	USA	USA	USA	China	USA	USA
2	Germany	Germany	Germany	Germany	Germany	Germany
3	China	Italy	France	USA	France	China
4	France	France	UK	Japan	Netherlands	France
5	Japan	Spain	Italy	France	China	Italy
6	UK	Belgium	Netherlands	Netherlands	Italy	Netherlands
7	Netherlands	Japan	Belgium	Italy	UK	Belgium
8	Italy	UK	Japan	Russian Federation	Belgium	UK
9	Belgium	Netherlands	China	UK	Spain	Japan
10	Canada	China	Spain	Belgium	Canada	Spain
11	Spain	Canada	Canada	Canada	India	Canada
12	Republic of Korea	Mexico	Russian Federation	Republic of Korea	Poland	Switzerland
13	Russian Federation	Republic of Korea	Republic of Korea	Mexico	Sweden	India
14	Mexico	Russian Federation	Switzerland	Saudi Arabia	Austria	Republic of Korea
15	Singapore	Poland	Austria	Singapore	Brazil	Poland
16	India	Austria	Poland	Spain	Australia	Turkey
17	Poland	Switzerland	Sweden	Malaysia	Japan	Czech Republic
18	Switzerland	Turkey	Mexico	Brazil	Russian Federation	Austria
19	Turkey	United Arab Emirates	India	India	Denmark	Thailand
20	Brazil	Denmark	Singapore	Switzerland	Thailand	Denmark

Table 2
Product names for SITC Rev. 1 3-digit code used in Fig. 6.

Symbol	Code	Abbreviation	Name
●	001	animals	Live animals
■	031	fish	Fish, fresh and simply preserved
♦	051	fruits	Fruit, fresh, and nuts excl. oil nuts
▲	054	vegetables	Vegetables, roots and tubers, fresh or dried
◀	061	sugarhon	Sugar and honey
▼	071	coffee	Coffee
▶	081	feedanim	Feed. stuff for animals excl. unmilled cereals
●	221	oilseeds	Oil seeds, oil nuts and oil kernels
■	263	cotton	Cotton
♦	283	ores	Ores and concentrates of non-ferrous base metals
●	331	petrolcrude	Petroleum, crude and partly refined
■	332	petrolprod	Petroleum products
♦	341	gas	Gas, natural and manufactured
●	512	orgchem	Organic chemicals
■	541	medpharm	Medicinal and pharmaceutical products
♦	581	plasticmat	Plastic materials, regenerated cellulose and resins
▲	599	chemmat	Chemical materials and products, n.e.s.
●	652	cottwoven	Cotton fabrics, woven ex. narrow or spec. fabrics
■	653	ncottwov	Textile fabrics, woven ex. narrow, spec., not cotton
♦	667	pearlsprec	Pearls and precious and semi precious stones
▲	674	iron	Universals, plates and sheets of iron or steel
◀	682	copper	Copper
●	711	nelecmach	Power generating machinery, other than electric
■	714	offmach	Office machines
♦	718	machindus	Machines for special industries
▲	719	mapplpart	Machinery and appliances non-electrical parts
◀	722	elecmach	Electric power machinery and switchgear
▼	724	telecomm	Telecommunications apparatus
▶	729	oelecmach	Other electrical machinery and apparatus
+	732	roadvehicles	Road motor vehicles
×	735	ships	Ships and boats
●	841	clothing	Clothing except fur clothing
●	931	finnotclass	Special transactions not class. accord. to kind

by a factor 5 (export) to 10 (import) smaller comparing to the random networks. This confirms the nestedness structure of WTN being similar to the case of ecology networks discussed in [12]. It is interesting to note that for random generated networks the values of T for import are larger than for export while to the WTN we have the opposite relation. The histogram of distribution of T for random generated networks for all years 1962–2009 is shown in Fig. S-5. Even minimal values of T remain several times larger than the WTN values of T.

In Fig. S-6 we show the dependence of T on the trade threshold μ for the WTN data in year 2008. We see that there is only about 10–20% of variation of T for the range $10^{-5} \leqslant \mu \leqslant 10^{-3}$. Even for a much larger range $10^{-6} \leqslant \mu \leqslant 10^{-2}$ the variation of T remains smooth and remains in the bounds of 100%. This confirms the stability of nestedness temperature in respect to broad range variations of μ. We present the majority of our data for $\mu = 10^{-3}$ which is approximately located in the flat range of T variation in year 2008. The data of Table 1 for EcoloRanking of countries at two different values of μ in year 2008 confirm the stability of this nestedness ordering. At the same time larger values of μ stress the importance of countries with a large trade volume, e.g. the position of China in export goes up from rank 5 at $\mu = 10^{-3}$ to rank 3 at $\mu = 10^{-2}$.

In Table 1 we present trade volume ranking and EcoloRanking of top 20 countries for import/export of WTN in year 2008.

In Table 2 we give the notations and symbols for Fig. 6 with corresponding SITC Rev. 1 codes and names. The list of all SITC Rev. 1 codes is available at [16] (see file http://unstats.un.org/unsd/tradekb/Attachment193.aspx). The colors of symbols in Fig. 4 mark the first digit of SITC Rev. 1 code: 0 – red (Food and live animals); 1 – does not appear in Fig. 4 (Beverages and tobacco); 2 – violet (Crude materials, inedible, except fuels); 3 – black (Mineral fuels, lubricants and related materials); 4 – does not appear in Fig. 4 (Animal and vegetable oils and fats); 5 – yellow (Chemicals); 6 – green (Manufactured goods classified chiefly by material); 7 – blue (Machinery and transport equipment); 8 – cyan (Miscellaneous manufactured articles); 9 – brown (Commod. and transacts. not class. accord. to kind).

References

[1] R.M. May, Stability and Complexity in Model Ecosystems, Princeton Univ. Press, New Jersey, USA, 2001.

[2] R.M. May, Nature 261 (1976) 459.

[3] E. Ott, Chaos in Dynamical Systems, Cambridge Univ. Press, Cambridge, UK, 2002.

[4] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of Networks, Oxford Univ. Press, Oxford, UK, 2003.

[5] G. Caldarelli, Scale-Free Networks, Oxford Univ. Press, Oxford, UK, 2007.

[6] G. Caldarelli, A. Vespignani (Eds.), Large Structure and Dynamics of Complex Networks, World Sci. Publ., Singapore, 2007.

[7] M. Pascual, J.A. Dunne (Eds.), Ecological Networks: Linking Structure to Dynamics in Food Webs, Oxford Univ. Press, Oxford, UK, 2006.

[8] J. Bascompte, P. Jordano, C.J. Melian, J.M. Olesen, Proc. Natl. Acad. Sci. USA 100 (2003) 9383.

[9] D.P. Vázquez, M.A. Aizen, Ecology 85 (2004) 1251.

[10] J. Memmott, N.M. Waser, M.V. Price, Proc. R. Soc. Lond. B 271 (2004) 2605.

[11] J.M. Olesen, J. Bascompte, Y.L. Dupont, P. Jordano, Proc. Natl. Acad. Sci. USA 104 (2007) 19891.

[12] E.L. Rezende, J.E. Lavabre, P.R. Guimarães, P. Jordano, J. Bascompte, Nature 448 (2007) 925.

[13] E. Burgos, H. Ceva, R.P.J. Perazzo, M. Devoto, D. Medan, M. Zimmermann, A.M. Delbue, J. Theor. Biol. 249 (2007) 307.

[14] U. Bastolla, M.A. Fortuna, A. Pascual-Garcia, A. Ferrera, B. Luque, J. Bascompte, Nature 458 (2009) 1018.

[15] S. Saaverda, D.B. Stouffer, B. Uzzi, J. Bascompte, Nature 478 (2011) 233.

[16] E. Burgos, H. Ceva, L. Hernández, R.P.J. Perazzo, M. Devoto, D. Medan, Phys. Rev. E 78 (2008) 046113;
E. Burgos, H. Ceva, L. Hernández, R.P.J. Perazzo, Comput. Phys. Commun. 180 (2009) 532.

[17] R.M. May, S.A. Levin, G. Sugihara, Nature 451 (2008) 893.

[18] A.G. Haldane, R.M. May, Nature 469 (2011) 351.

[19] United Nations Commodity Trade Statistics Database, http://comtrade.un.org/db/.

[20] M.A. Rodríguez-Gironés, L. Santamaría, J. Biogeogr. 33 (2006) 924.

[21] http://ieg.ebd.csic.es/JordiBascompte/Resources.html.

[22] Central Intelligence Agency, The CIA Wold Factbook 2010, Skyhorse Publ. Inc., 2009.

[23] L. Ermann, D.L. Shepelyansky, Acta Phys. Pol. A 120 (2011) A-158, http://www.quantware.ups-tlse.fr/QWLIB/tradecheirank/.

Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles

Young-Ho Eom, Dima L. Shepelyansky*

Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, Toulouse, France

Abstract

How different cultures evaluate a person? Is an important person in one culture is also important in the other culture? We address these questions via ranking of multilingual Wikipedia articles. With three ranking algorithms based on network structure of Wikipedia, we assign ranking to all articles in 9 multilingual editions of Wikipedia and investigate general ranking structure of PageRank, CheiRank and 2DRank. In particular, we focus on articles related to persons, identify top 30 persons for each rank among different editions and analyze distinctions of their distributions over activity fields such as politics, art, science, religion, sport for each edition. We find that local heroes are dominant but also global heroes exist and create an effective network representing entanglement of cultures. The Google matrix analysis of network of cultures shows signs of the Zipf law distribution. This approach allows to examine diversity and shared characteristics of knowledge organization between cultures. The developed computational, data driven approach highlights cultural interconnections in a new perspective. Dated: June 26, 2013

Citation: Eom Y-H, Shepelyansky DL (2013) Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles. PLoS ONE 8(10): e74554. doi:10.1371/journal.pone.0074554

Editor: Matjaz Perc, University of Maribor, Slovenia

Received June 26, 2013; **Accepted** August 5, 2013; **Published** October 3, 2013

Funding: This research is supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE number 288956). No additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dima@irsamc.ups-tlse.fr

Introduction

Wikipedia, the online collaborative encyclopedia, is an amazing example of human collaboration for knowledge description, characterization and creation. Like the Library of Babel, described by Jorge Luis Borges [1], Wikipedia goes to accumulate the whole human knowledge. Since every behavioral 'footprint' (log) is recorded and open to anyone, Wikipedia provides great opportunity to study various types of social aspects such as opinion consensus [2,3], language complexity [4], and collaboration structure [5–7]. A remarkable feature of Wikipedia is its existence in various language editions. In a first approximation we can attribute each language to an independent culture, leaving for future refinements of cultures inside one language. Although Wikipedia has a neutral point of view policy, cultural bias or reflected cultural diversity is inevitable since knowledge and knowledge description are also affected by culture like other human behaviors [8–11]. Thus the cultural bias of contents [12] becomes an important issue. Similarity features between various Wikipedia editions has been discussed at [13]. However, the cross-cultural difference between Wikipedia editions can be also a valuable opportunity for a cross-cultural empirical study with quantitative approach. Recent steps in this direction, done for biographical networks of Wikipedia, have been reported in [14].

Here we address the question of how importance (ranking) of an article in Wikipedia depends on cultural diversity. In particular, we consider articles about persons. For instance, is an important person in English Wikipedia is also important in Korean Wikipedia? How about French? Since Wikipedia is the product of collective intelligence, the ranking of articles about persons is a collective evaluation of the persons by Wikipedia users. For the ranking of Wikipedia articles we use PageRank algorithm of Brin and Page [15], CheiRank and 2Drank algorithms used in [16–18], which allow to characterize the information flows with incoming and outgoing links. We also analyze the distribution of top ranked persons over main human activities attributed to politics, science, art, religion, sport, etc (all others), extending the approach developed in [17,19] to multiple cultures (languages). The comparison of different cultures shows that they have distinct dominance of these activities.

We attribute belongings of top ranked persons at each Wikipedia language to different cultures (native languages) and in this way construct the network of cultures. The Google matrix analysis of this network allows us to find interconnections and entanglement of cultures. We believe that our computational and statistical analysis of large-scale Wikipedia networks, combined with comparative distinctions of different languages, generates novel insights on cultural diversity.

Methods

We consider Wikipedia as a network of articles. Each article corresponds to a node of the network and hyperlinks between articles correspond to links of the network. For a given network, we can define adjacency matrix A_{ij}. If there is a link (one or more quotations) from node (article) j to node (article) i then $A_{ij}=1$, otherwise, $A_{ij}=0$. The out-degree $k_{out}(j)$ is the number of links from node j to other nodes and the in-degree $k_{in}(j)$ is the number of links to node j from other nodes.

Google matrix

The matrix S_{ij} of Markov chain transitions is constructed from adjacency matrix A_{ij} by normalizing sum of elements of each column to unity ($S_{ij}=A_{ij}/\sum_i A_{ij}$, $\sum_i S_{ij}=1$) and replacing columns with only zero elements (*dangling nodes*) by $1/N$, with N being the matrix size. Then the Google matrix of this directed network has the form [15,20]:

$$G_{ij}=\alpha S_{ij}+(1-\alpha)/N. \tag{1}$$

In the WWW context the damping parameter α describes the probability $(1-\alpha)$ to jump to any article (node) for a random walker. The matrix G belongs to the class of Perron-Frobenius operators, it naturally appears in dynamical systems [21]. The right eigenvector at $\lambda=1$, which is called the PageRank, has real non-negative elements $P(i)$ and gives a probability $P(i)$ to find a random walker at site i. It is possible to rank all nodes in a decreasing order of PageRank probability $P(K(i))$ so that the PageRank index $K(i)$ sorts all N nodes i according their ranks. For large size networks the PageRank vector and several other eigenvectors can be numerically obtained using the powerful Arnoldi algorithm as described in [22]. The PageRank vector can be also obtained by a simple iteration method [20]. Here, we use here the standard value of $\alpha=0.85$ [20].

To rank articles of Wikipedia, we use three ranking algorithms based on network structure of Wikipedia articles. Detail description of these algorithms and their use for English Wikipedia articles are given in [17–19,22].

PageRank algorithm

PageRank algorithm is originally introduced for Google web search engine to rank web pages of the World Wide Web (WWW) [15]. Currently PageRank is widely used to rank nodes of network systems including scientific papers [23], social network services [24] and even biological systems [25]. Here we briefly outline the iteration method of PageRank computation. The PageRank vector $P(i,t)$ of a node i at iteration t in a network of N nodes is given by

$$\begin{aligned} P(i,t)&=\sum_j G_{ij}\ P(j,t-1)\ ,\ P(i,t)\\ &=(1-\alpha)/N+\alpha\sum_j A_{ij}\ P(j,t-1)/k_{out}(j). \end{aligned} \tag{2}$$

The stationary state $P(i)$ of $P(i,t)$ is the PageRank of node i. More detail information about PageRank algorithm is described in [20]. Ordering all nodes by their decreasing probability $P(i)$ we obtain the PageRank index $K(i)$.

The essential idea of PageRank algorithm is to use a directed link as a weighted 'recommendation'. Like in academic citation network, more cited nodes are considered to be more important. In addition, recommendations by highly ranked articles are more important. Therefore high PageRank nodes in the network have many incoming links from other nodes or incoming links from high PageRank nodes.

CheiRank algorithm

While the PageRank algorithm uses information of incoming links to node i, CheiRank algorithm considers information of outgoing links from node i [16–18]. Thus CheiRank is complementary to PageRank in order to rank nodes in directed networks. The CheiRank vector $P^*(i,t)$ of a node at iteration time t is given by

$$P^*(i)=(1-\alpha)/N+\alpha\sum_j A_{ji}\ P^*(j)/k_{in}(j) \tag{3}$$

We also point out that the CheiRank is the right eigenvector with maximal eigenvalue $\lambda=1$ satisfying the equation $P^*(i)=\sum_j G^*_{ij}\ P^*(j)$, where the Google matrix G^* is built for the network with inverted directions of links via the standard definition of G given above.

Table 1. Considered Wikipedia networks from language editions: English (EN), French (FR), German (DE), Italian (IT), Spanish (ES), Dutch (NL), Russian (RU), Hungarian (HU), Korean (KO).

Edition	N_A	N_L	κ	Date
EN	3920628	92878869	3.905562	Mar. 2012
FR	1224791	30717338	3.411864	Feb. 2012
DE	1396293	32932343	3.342059	Mar. 2012
IT	917626	22715046	7.953106	Mar. 2012
ES	873149	20410260	3.443931	Feb. 2012
NL	1034912	14642629	7.801457	Feb. 2012
RU	830898	17737815	2.881896	Feb. 2012
HU	217520	5067189	2.638393	Feb. 2012
KO	323461	4209691	1.084982	Feb. 2012

Here N_A is number of articles, N_L is number of hyperlinks between articles, κ is the correlator between PageRank and CheiRank. Date represents the time in which data are collected.

doi:10.1371/journal.pone.0074554.t001

Like for PageRank, we consider the stationary state $P^*(i)$ of $P^*(i,t)$ as the CheiRank probability of node i at $\alpha=0.85$. High CheiRank nodes in the network have a large out-degree. Ordering all nodes by their decreasing probability $P^*(i)$ we obtain the CheiRank index $K^*(i)$.

We note that PageRank and CheiRank naturally appear in the world trade network corresponding to import and export in a commercial exchange between countries [26].

The correlation between PageRank and CheiRank vectors can be characterized by the correlator κ [16–18] defined by

$$\kappa=N\sum_i P(i)\ P^*(i)-1 \tag{4}$$

The value of correlator for each Wikipedia edition is represented in Table 1. All correlators are positive and distributed in the interval (1,8).

2DRank algorithm

With PageRank $P(i)$ and CheiRank $P^*(i)$ probabilities, we can assign PageRank ranking $K(i)$ and CheiRank ranking $K^*(i)$ to each article, respectively. From these two ranks, we can construct 2-dimensional plane of K and K^*. The two dimensional ranking K_2 is defined by counting nodes in order of their appearance on ribs of squares in (K,K^*) plane with the square size growing from $K=1$ to $K=N$ [17]. A direct detailed illustration and description of this algorithm is given in [17]. Briefly, nodes with high PageRank and CheiRank both get high 2DRank ranking.

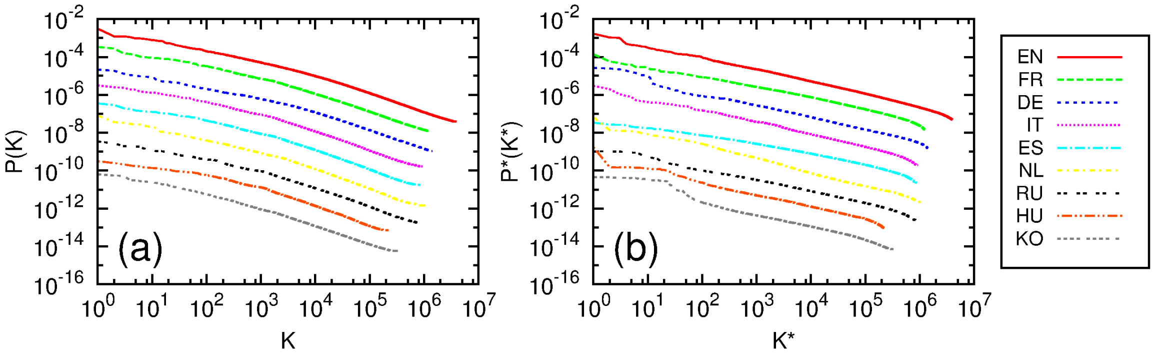


Figure 1. PageRank probability $P(K)$ as function of PageRank index K (a) and CheiRank probability $P^*(K^*)$ as function of CheiRank index K^* (b). For a better visualization each PageRank P and CheiRank P^* curve is shifted down by a factor 10^0 (EN), 10^1 (FR), 10^2 (DE), 10^3 (IT), 10^4 (ES), 10^5 (NL), 10^6 (RU), 10^7 (HU), 10^8 (KO).
doi:10.1371/journal.pone.0074554.g001

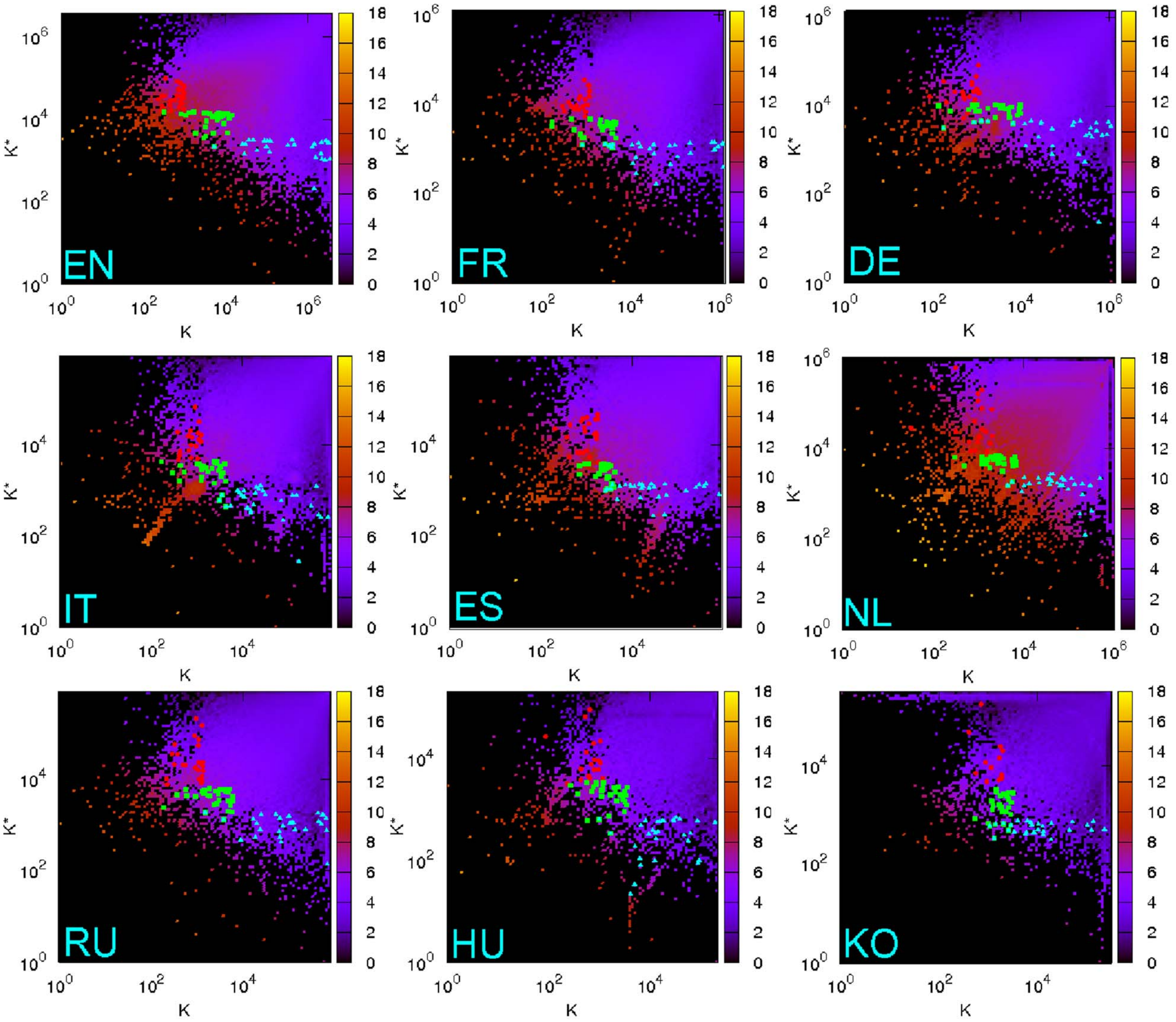


Figure 2. Density of Wikipedia articles in the PageRank ranking K versus CheiRank ranking K^* plane for each Wikipedia edition. The red points are top PageRank articles of persons, the green points are top 2DRank articles of persons and the cyan points are top CheiRank articles of persons. Panels show: English (top-left), French (top-center), German (top-right), Italian (middle-left), Spanish (middle-center), Dutch (middle-left), Russian (bottom-left), Hungarian (bottom-center), Korean (bottom-right). Color bars shown natural logarithm of density, changing from minimal nonzero density (dark) to maximal one (white), zero density is shown by black.
doi:10.1371/journal.pone.0074554.g002

Table 2. Example of list of top 10 persons by PageRank for English Wikipedia with their field of activity and native language.

$R_{EN,PageRank}$	Person	Field	Culture	Locality
1	Napoleon	Politics	FR	Non-local
2	Carl Linnaeus	Science	WR	Non-local
3	George W. Bush	Politics	EN	Local
4	Barack Obama	Politics	EN	Local
5	Elizabeth II	Politics	EN	Local
6	Jesus	Religion	WR	Non-local
7	William Shakespeare	Art	EN	Local
8	Aristotle	Science	WR	Non-local
9	Adolf Hitler	Politics	DE	Non-local
10	Bill Clinton	Politics	EN	Local

doi:10.1371/journal.pone.0074554.t002

Data Description

We consider 9 editions of Wikipedia including English (EN), French (FR), German (DE), Italian (IT), Spanish (ES), Dutch (NL), Russian (RU), Hungarian (HU) and Korean (KO). Since Wikipedia has various language editions and language is a most fundamental part of culture, the cross-edition study of Wikipedia can give us insight on cultural diversity. The overview summary of parameters of each Wikipedia is represented in Table 1.

The corresponding networks of these 9 editions are collected and kindly provided to us by S.Vigna from LAW, Univ. of Milano. The first 7 editions in the above list represent mostly spoken European languages (except Polish). Hungarian and Korean are additional editions representing languages of not very large population on European and Asian scales respectively. They allow us to see interactions not only between large cultures but also to see links on a small scale. The KO and RU editions allow us to compare views from European and Asian continents. We also note that in part these 9 editions reflect the languages present in the EC NADINE collaboration.

We understand that the present selection of Wikipedia editions does represent a complete view of all 250 languages present at Wikipedia. However, we think that this selection allows us to perform the quantitative statistical analysis of interactions between cultures making a first step in this direction.

To analyze these interactions we select the fist top 30 persons (or articles about persons) appearing in the top ranking list of each of 9 editions for 3 ranking algorithms of PageRank, CheiRank and 2DRank. We select these 30 persons manually analyzing each list. We attribute each of 30 persons to one of 6 fields of human activity: politics, science, art, religion, sport, and etc (here "etc" includes all other activities). In addition we attribute each person to one of 9 selected languages or cultures. We place persons belonging to other languages inside the additional culture WR (world) (e.g. Plato). Usually a belonging of a person to activity field

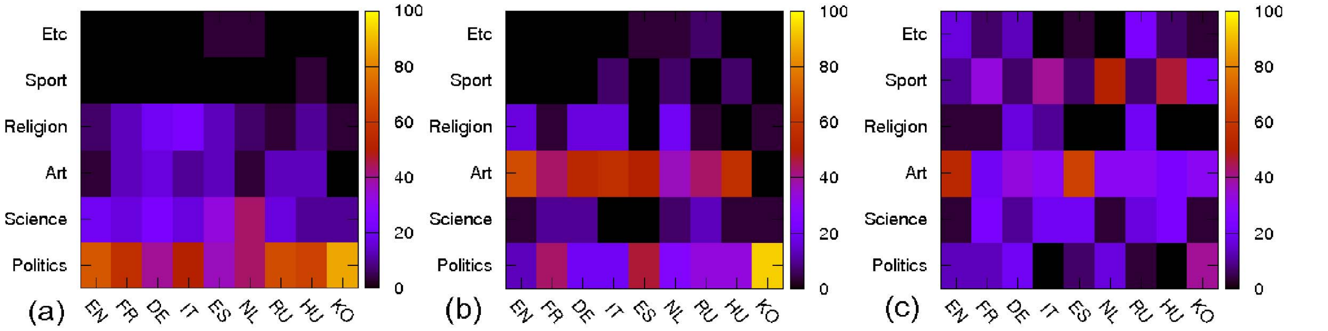


Figure 3. Distribution of top 30 persons in each rank over activity fields for each Wikipedia edition. Panels correspond to (a) PageRank, (b) 2DRank, (3) CheiRank. The color bar shows the values in percents.
doi:10.1371/journal.pone.0074554.g003

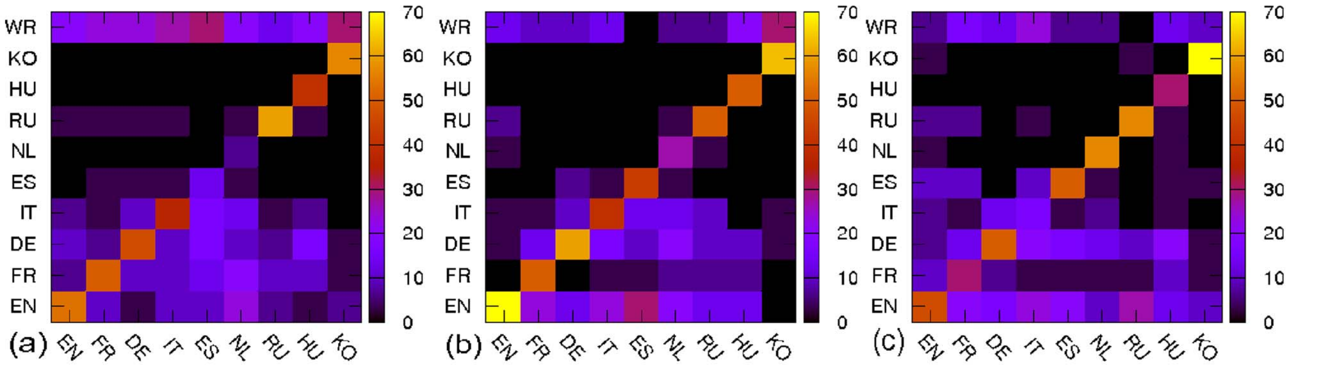


Figure 4. Distributions of top 30 persons over different cultures corresponding to Wikipedia editions, "WR" category represents all other cultures which do not belong to considered 9 Wikipedia editions. Panels show ranking by (a) PageRank, (b) 2DRank, (3) CheiRank. The color bar shows the values in percents.
doi:10.1371/journal.pone.0074554.g004

Table 3. PageRank contribution per link and in-degree of PageRank local and non-local heroes i for each edition.

Edition	N_{Local}	$[P(j)/k(j)_{out}]_L$		$[P(j)/k(j)_{out}]_{NL}$	$[k(L)_{in}]$		$[k(NL)_{in}]$
EN	16	1.43×10^{-8}	$<$	2.18×10^{-8}	5.3×10^{3}	$>$	3.1×10^{3}
FR	15	3.88×10^{-8}	$<$	5.69×10^{-8}	2.6×10^{3}	$>$	2.0×10^{3}
DE	14	3.48×10^{-8}	$<$	4.29×10^{-8}	2.6×10^{3}	$>$	2.1×10^{3}
IT	11	7.00×10^{-8}	$<$	7.21×10^{-8}	1.9×10^{3}	$>$	1.5×10^{3}
ES	4	5.44×10^{-8}	$<$	8.58×10^{-8}	2.2×10^{3}	$>$	1.2×10^{3}
NL	2	7.77×10^{-8}	$<$	14.4×10^{-8}	1.0×10^{3}	$>$	6.7×10^{2}
RU	18	6.67×10^{-8}	$<$	10.2×10^{-8}	1.7×10^{3}	$>$	1.5×10^{3}
HU	12	21.1×10^{-8}	$<$	32.3×10^{-8}	8.1×10^{2}	$>$	5.3×10^{2}
KO	17	16.6×10^{-8}	$<$	35.5×10^{-8}	4.7×10^{2}	$>$	2.3×10^{2}

$[P(j)/k(j)_{out}]_L$ and $[P(j)/k(j)_{out}]_{NL}$ are median PageRank contribution of a local hero L and non-local hero NL by a article j which cites local heroes L and non-local heroes NL respectively. $[k(L)_{in}]$ and $[k(NL)_{in}]$ are median number of in-degree $k(L)_{in}$ and $k(NL)_{in}$ of local hero L and non-local hero NL, respectively. N_{Local} is number local heroes in given edition.
doi:10.1371/journal.pone.0074554.t003

and language is taken from the English Wikipedia article about this person. If there is no such English Wikipedia article then we use an article of a Wikipedia edition language which is native for such a person. Usually there is no ambiguity in the distribution over activities and languages. Thus Christopher Columbus is attributed to IT culture and activity field etc, since English Wikipedia describes him as "italian explorer, navigator, and colonizer". By our definition politics includes politicians (e.g. Barak Obama), emperors (e.g. Julius Caesar), kings (e.g. Charlemagne). Arts includes writers (e.g. William Shakespeare), singers (e.g. Frank Sinatra), painters (Leonardo da Vinci), architects, artists, film makers (e.g. Steven Spielberg). Science includes physicists, philosophers (e.g. Plato), biologists, mathematicians and others. Religion includes such persons as Jesus, Pope John Paul II. Sport includes sportsmen (e.g. Roger Federer). All other activities are placed in activity etc (e.g. Christopher Columbus, Yuri Gagarin). Each person belongs only to one language and one activity field. There are only a few cases which can be questioned, e.g. Charles V, Holy Roman Emperor who is attributed to ES language since from early long times he was the king of Spain. All listings of person distributions over the above categories are presented at the web page given at Supporting Information (SI) file and in 27 tables given in File S1.

Unfortunately, we were obliged to construct these distributions manually following each person individually at the Wikipedia ranking listings. Due to that we restricted our analysis only to top 30 persons. We think that this number is sufficiently large so that the statistical fluctuations do not generate significant changes. Indeed, we find that our EN distribution over field activities is close to the one obtained for 100 top persons of English Wikipedia dated by Aug 2009 [17].

To perform additional tests we use the database of about 250000 person names in English, Italian and Dutch from the research work [14] provided to us by P.Aragón and A.Kaltenbrunner. Using this database we were able to use computerized (automatic) selection of top 100 persons from the ranking lists and to compare their distributions over activities and languages with our case of 30 persons. The comparison is presented in figures S1,S2,S3 in File S1. For these 3 cultures we find that our top 30 persons data are statistically stable even if the fluctuations are larger for CheiRank lists. This is in an agreement with the fact that the CheiRank probabilities. related to the outgoing links, are more fluctuating (see discussion at [19]).

Of course, it would be interesting to extend the computerized analysis of personalities to a larger number of top persons and larger number of languages. However, the database of persons in various languages still should be cleaned and checked and also attribution of persons to various activities and languages still requires a significant amount of work. Due to that we present here our analysis only for 30 top persons. But we note that by itself it represents an interesting case study since here we have the most important persons for each ranking. May be the top 1000 persons would be statistically more stable but clearly a person at position 30 is more important than a one at position 1000. Thus we think that the top 30 persons already give an interesting information on links and interactions between cultures. This information can be used in future more extended studies of a larger number of persons and languages.

Finally we note that the language is the primary element of culture even if, of course, culture is not reduced only to language. In this analysis we use in a first approximation an equivalence between language and culture leaving for future studies the refinement of this link which is of course much more complex. In this approximation we consider that a person like Mahatma Gandhi belongs to EN culture since English is the official language of India. A more advanced study should take into account Hindi

Table 4. List of local heroes by PageRank for each Wikipedia edition.

Edition	1st	2nd	3rd
EN	George W. Bush	Barack Obama	Elizabeth II
FR	Napoleon	Louis XIV of France	Charles de Gaulle
DE	Adolf Hitler	Martin Luther	Immanuel Kant
IT	Augustus	Dante Alighieri	Julius Caesar
ES	Charles V, Holy Roman Emperor	Philip II of Spain	Francisco Franco
NL	William I of the Netherlands	Beatrix of the Netherlands	William the Silent
RU	Peter the Great	Joseph Stalin	Alexander Pushkin
HU	Matthias Corvinus	Szentágothai János	Stephen I of Hungary
KO	Gojong of the Korean Empire	Sejong the Great	Park Chung-hee

All names are represented by article titles in English Wikipedia. Here "William the Silent" is the third local hero in Dutch Wikipedia but he is out of top 30 persons.
doi:10.1371/journal.pone.0074554.t004

Table 5. List of local heroes by CheiRank for each Wikipedia edition.

Edition	1st	2nd	3rd
EN	C. H. Vijayashankar	Matt Kelley	William Shakespeare (inventor)
FR	Jacques Davy Duperron	Jean Baptiste Eblé	Marie-Magdeleine Aymé de La Chevrelière
DE	Harry Pepl	Marc Zwiebler	Eugen Richter
IT	Nduccio	Vincenzo Olivieri	Mina (singer)
ES	Che Guevara	Arturo Mercado	Francisco Goya
NL	Hans Renders	Julian Jenner	Marten Toonder
RU	Aleksander Vladimirovich Sotnik	Aleksei Aleksandrovich Bobrinsky	Boris Grebenshchikov
HU	Csernus Imre	Kati Kovács	Pléh Csaba
KO	Lee Jong-wook (baseball)	Kim Dae-jung	Kim Kyu-sik

All names are represented by article titles in English Wikipedia.
doi:10.1371/journal.pone.0074554.t005

Table 6. List of local heroes by 2DRank for each Wikipedia edition.

Edition	1st	2nd	3rd
EN	Frank Sinatra	Paul McCartney	Michael Jackson
FR	François Mitterrand	Jacques Chirac	Honoré de Balzac
DE	Adolf Hitler	Otto von Bismarck	Ludwig van Beethoven
IT	Giusppe Garibaldi	Raphael	Benito Mussolini
ES	Simón Bolívar	Francisco Goya	Fidel Castro
NL	Albert II of Belgium	Johan Cruyff	Rembrandt
RU	Dmitri Mendeleev	Peter the Great	Yaroslav the Wise
HU	Stephen I of Hungary	Sándor Petöfi	Franz Liszt
KO	Gojong of the Korean Empire	Sejong the Great	Park Chung-hee

All names are represented by article titles in English Wikipedia.
doi:10.1371/journal.pone.0074554.t006

Wikipedia edition and attribute this person to this edition. Definitely our statistical study is only a first step in Wikipedia based statistical analysis of network of cultures and their interactions.

We note that any person from our top 30 ranking belongs only to one activity field and one culture. We also define local heros as those who in a given language edition are attributed to this language, and non-local heros as those who belong in a given edition to other languages. We use category WR (world) where we

Table 7. List of global heroes by PageRank and 2DRank for all 9 Wikipedia editions.

Rank	PageRank global heroes	Θ_{PR}	N_A	2DRank global heroes	Θ_{2D}	N_A
1st	Napoleon	259	9	Micheal Jackson	119	5
2nd	Jesus	239	9	Adolf Hitler	93	6
3rd	Carl Linnaeus	235	8	Julius Caesar	85	5
4th	Aristotle	228	9	Pope Benedict XVI	80	4
5th	Adolf Hitler	200	9	Wolfgang Amadeus Mozart	75	5
6th	Julius Caesar	161	8	Pope John Paul II	71	4
7th	Plato	119	6	Ludwig van Beethoven	69	4
8th	Charlemagne	111	8	Bob Dylan	66	4
9th	William Shakespeare	110	7	William Shakespeare	57	3
10th	Pope John Paul II	108	6	Alexander the Great	56	3

All names are represented by article titles in English Wikipedia. Here, Θ_A is the ranking score of the algorithm A (5); N_A is the number of appearances of a given person in the top 30 rank for all editions.
doi:10.1371/journal.pone.0074554.t007

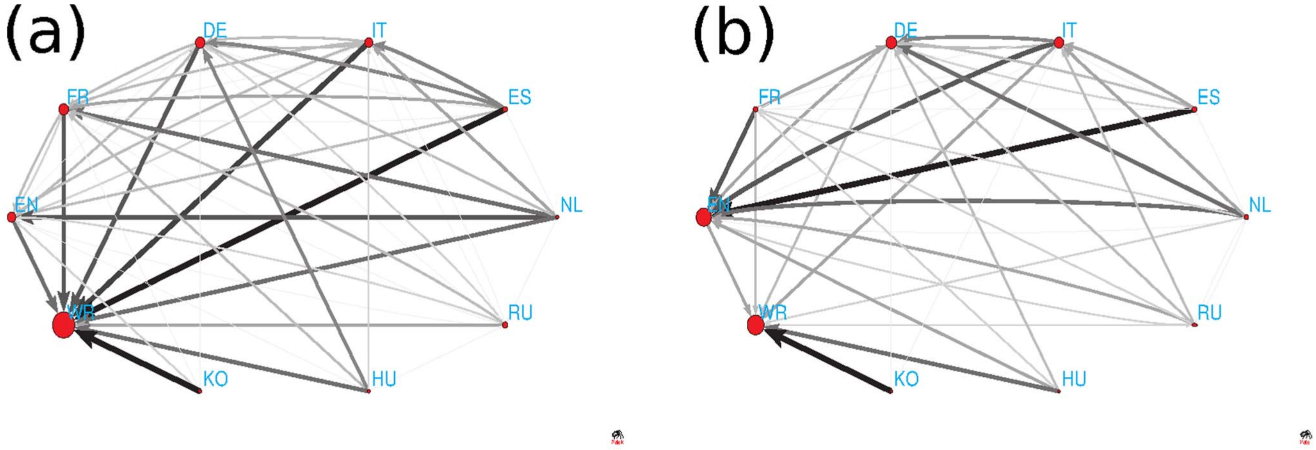


Figure 5. Network of cultures obtained from 9 Wikipedia languages and the remaining world (WR) selecting 30 top persons of PageRank (a) and 2DRank (b) in each culture. The link width and darkness are proportional to a number of foreign persons quoted in top 30 of a given culture, the link direction goes from a given culture to cultures of quoted foreign persons, quotations inside cultures are not considered. The size of nodes is proportional to their PageRank.
doi:10.1371/journal.pone.0074554.g005

place persons who do not belong to any of our 9 languages (e.g. Pope John Paul II belongs to WR since his native language is Polish).

Results

We investigate ranking structure of articles and identify global properties of PageRank and CheiRank vectors. The detailed analysis is done for top 30 persons obtained from the global list of ranked articles for each of 9 languages. The distinctions and common characteristics of cultures are analyzed by attributing top 30 persons in each language to human activities listed above and to their native language.

General ranking structure

We calculate PageRank and CheiRank probabilities and indexes for all networks of considered Wikipedia editions. The PageRank and CheiRank probabilities as functions of ranking indexes are shown in Fig. 1. The decay is compatible with an approximate algebraic decrease of a type $P \sim 1/K^{\beta}$, $P^* \sim 1/K^{*\beta}$ with $\beta \sim 1$ for PageRank and $\beta \sim 0.6$ for CheiRank. These values are similar to those found for the English Wikipedia of 2009 [17]. The difference of β values originates from asymmetric nature between in-degree and out-degree distributions, since PageRank is based on incoming edges while CheiRank is based on outgoing edges. In-degree distribution of Wikipedia editions is broader than out-degree distribution of the same edition. Indeed, the CheiRank probability is proportional to frequency of outgoing links which has a more rapid decay compared to incoming one (see discussion in [17]). The PageRank (CheiRank) probability distributions are similar for all editions. However, the fluctuations of P^* are stronger that is related to stronger fluctuations of outgoing edges [19].

The top article of PageRank is usually *USA* or the name of country of a given language (FR, RU, KO). For NL we have at the top *beetle, species, France*. The top articles of CheiRank are various listings.

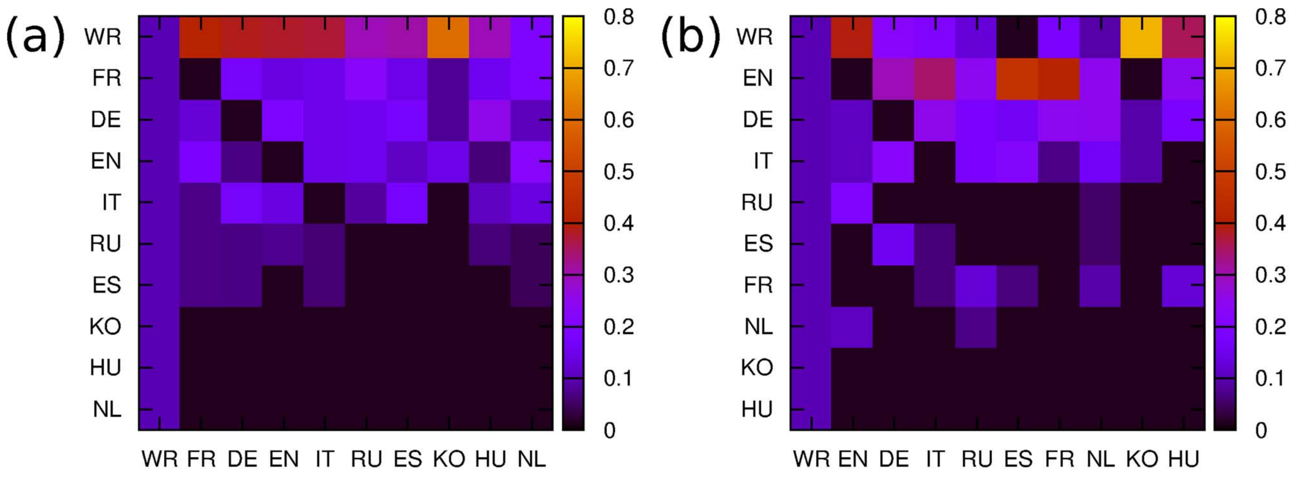


Figure 6. Google matrix of network of cultures from Fig. 5, shown respectively for panels (a),(b)**.** The matrix elements G_{ij} are shown by color at the damping factor $\alpha = 0.85$, index j is chosen as the PageRank index K of PageRank vector so that the top cultures with $K = K' = 1$ are located at the top left corner of the matrix.
doi:10.1371/journal.pone.0074554.g006

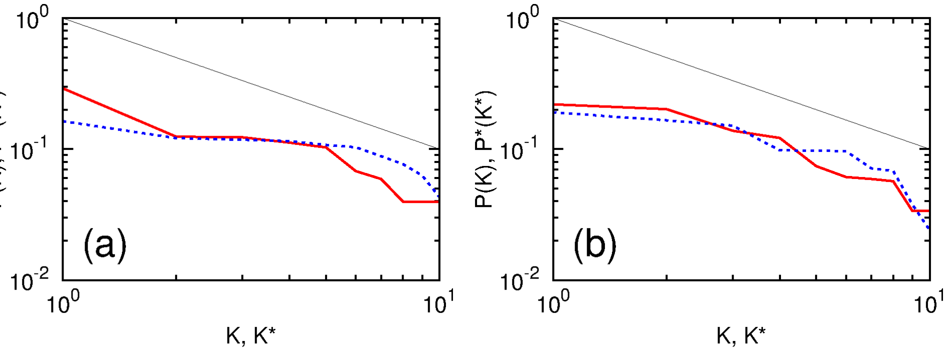


Figure 7. Dependence of probabilities of PageRank P (red) and CheiRank P^* (blue) on corresponding indexes K and K^*. The probabilities are obtained from the network and Google matrix of cultures shown in Fig. 5 and Fig. 6 for corresponding panels (a),(b). The straight lines indicate the Zipf law $P \sim 1/K$; $P^* \sim 1/K^*$.
doi:10.1371/journal.pone.0074554.g007

Since each article has its PageRank ranking K and CheiRank ranking K^*, we can assign two dimensional coordinates to all the articles. Fig. 2 shows the density of articles in the two dimensional plane (K,K^*) for each Wikipedia edition. The density is computed for 100×100 logarithmically equidistant cells which cover the whole plane (K,K^*). The density plot represents the locations of articles in the plane. We can observe high density of articles around line $K = K^* + const$ that indicates the positive correlation between PageRank and CheiRank. However, there are only a few articles within the region of top both PageRank and CheiRank indexes. We also observe the tendency that while high PageRank articles $(K < 100)$ have intermediate CheiRank $(10^2 < K^* < 10^4)$, high CheiRank articles $(K^* < 100)$ have broad PageRank rank values.

Ranking of articles for persons

We choose top 30 articles about persons for each edition and each ranking. In Fig. 2, they are shown by red circles (PageRank), green squares (2DRank) and cyan triangles (CheiRank). We assign local ranking $R_{E,A}$ $(1\ldots30)$ to each person in the list of top 30 persons for each edition E and ranking algorithm A. An example of $E = EN$ and $A = PageRank$ are given in Table 2.

From the lists of top persons, we identify the "fields" of activity for each top 30 rank person in which he/she is active on. We categorize six activity fields - politics, art, science, religion, sport and etc (here "etc" includes all other activities). As shown in Fig. 3, for PageRank, politics is dominant and science is secondarily dominant. The only exception is Dutch where science is the almost dominant activity field (politics has the same number of points). In case of 2DRank, art becomes dominant and politics is secondarily dominant. In case of CheiRank, art and sport are dominant fields. Thus for example, in CheiRank top 30 list we find astronomers who discovered a lot of asteroids, e.g. Karl Wilhelm Reinmuth (4th position in RU and 7th in DE), who was a prolific discoverer of about 400 of them. As a result, his article contains a long listing of asteroids discovered by him giving him a high CheiRank.

The change of activity priority for different ranks is due to the different balance between incoming and outgoing links there. Usually the politicians are well known for a broad public, hence, the articles about politicians are pointed by many articles. However, the articles about politician are not very communicative since they rarely point to other articles. In contrast, articles about persons in other fields like science, art and sport are more communicative because of listings of insects, planets, asteroids they discovered, or listings of song albums or sport competitions they gain.

Next we investigate distributions over "cultures" to which persons belong. We determined the culture of person based on the language the person mainly used (mainly native language). We consider 10 culture categories - EN, FR, DE, IT, ES, NL, RU, HU, KO and WR. Here "WR" category represents all other cultures which do not belong to considered 9 Wikipedia editions.

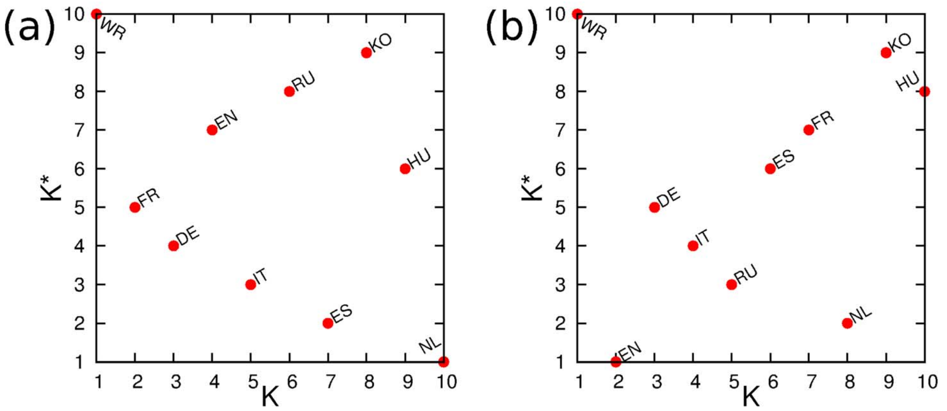


Figure 8. PageRank versus CheiRank plane of cultures with corresponding indexes K and K^* obtained from the network of cultures for corresponding panels (a),(b).
doi:10.1371/journal.pone.0074554.g008

Comparing with the culture of persons at various editions, we can assign “locality” to each 30 top rank persons for a given Wikipedia edition and ranking algorithm. For example, as shown in Table 2, *George W. Bush* belongs to “Politics”, “English” and “Local” for English Wikipedia and PageRank, while *Jesus* belongs to “Religion”, “World” WR and “Non-local”.

As shown in Fig. 4, regardless of ranking algorithms, main part of top 30 ranking persons of each edition belong to the culture of the edition (usually about 50%). For example, high PageRank persons in English Wikipedia are mainly English (53.3%). This corresponds to the self-focusing effect discussed in [6]. It is notable that top ranking persons in Korean Wikipedia are not only mainly Korean (56.7%) but also the most top ranking non Korean persons in Korean Wikipedia are Chinese and Japanese (20%). Although there is a strong tendency that each edition favors its own persons, there is also overlap between editions. For PageRank, on average, 23.7 percent of top persons are overlapping while for CheiRank , the overlap is quite low, only 1.3 percent. For 2DRank, the overlap is 6.3 percent. The overlap of list of top persons implies the existence of cross-cultural ‘heroes’.

To understand the difference between local and non-local top persons for each edition quantitatively, we consider the PageRank case because it has a large fraction of non-local top persons. From Eq. (2), a citing article j contributes $\langle P(j)/k_{out}(j)\rangle$ to PageRank of a node i. So the PageRank $P(i)$ can be high if the node i has many incoming links from citing articles j or it has incoming links from high PageRank nodes j with low out-degree $k_{out}(j)$. Thus we can identify origin of each top person's PageRank using the average PageRank contribution $\langle P(j)/k_{out}(j)\rangle$ by nodes j to person i and average number of incoming edges (in-degree) $k_{in}(i)$ of person i .

As represented in Table 3, considering median, local top persons have more incoming links than non-local top persons but the PageRank contribution of the corresponding links are lower than links of non-local top persons. This indicates that local top persons are cited more than non-local top persons but non-local top persons are cited more high weighted links (i.e. cited by important articles or by articles which don't have many citing links).

Global and local heroes

Based on cultural dependency on rankings of persons, we can identify global and local heroes in the considered Wikipedia editions. However, for CheiRank the overlap is very low and our statistics is not sufficient for selection of global heroes. Hence we consider only PageRank and 2DRank cases. We determine the local heroes for each ranking and for each edition as top persons of the given ranking who belongs to the same culture as the edition. Top 3 local heroes for each ranking and each edition are represented in Table 4 (PageRank), Table 5 (CheiRank) and Table 6 (2DRank), respectively.

In order to identify the global heroes, we define ranking score $\Theta_{P,A}$ for each person P and each ranking algorithm A. Since every person in the top person list has relative ranking $R_{P,E,A}$ for each Wikipedia edition E and ranking algorithm A (For instance, in Table 2, $R_{Napoleon,EN,PageRank}=1$). The ranking score $\Theta_{P,A}$ of a person P is give by

$$\Theta_{P,A}=\sum_{E}(31-R_{P,E,A}) \qquad (5)$$

According to this definition, a person who appears more often in the lists of editions and has top ranking in the list gets high ranking score. We sort this ranking score for each algorithm. In this way obtain a list of global heroes for each algorithm. The result is shown in Table 7. Napoleon is the 1st global hero by PageRank and Micheal Jackson is the 1st global hero by 2DRank.

Network of cultures

To characterize the entanglement and interlinking of cultures we use the data of Fig. 4 and from them construct the network of cultures. The image of networks obtained from top 30 persons of PageRank and 2DRank listings are shown in Fig. 5 (we do not consider CheiRank case due to small overlap of persons resulting in a small data statistics). The weight of directed Markov transition, or number of links, from a culture A to a culture B is given by a number of persons of a given culture B (e.g FR) appearing in the list of top 30 persons of PageRank (or 2DRank) in a given culture A (e.g. EN). Thus e.g. for transition from EN to FR in PageRank we find 2 links (2 French persons in PageRank top 30 persons of English Wikipedia); for transition from FR to EN in PageRank we have 3 links (3 English persons in PageRank top 30 persons of French Wikipedia). The transitions inside each culture (persons of the same language as language edition) are omitted since we are analyzing the interlinks between cultures. Then the Google matrix of cultures is constructed by the standard rule for the directed networks: all links are treated democratically with the same weight, sum of links in each column is renormalized to unity, $\alpha=0.85$. Even if this network has only 10 nodes we still can find for it PageRank and CheiRank probabilities P and P^{*} and corresponding indexes K and K^{*}. The matrix elements of G matrix, written in order of index K, are shown in Fig. 6 for the corresponding networks of cultures presented in Fig. 5. We note that we consider all cultures on equal democratic grounds.

The decays of PageRank and CheiRank probabilities with the indexes K,K^{*} are shown in Fig. 7 for the culture networks of Fig. 5. On a first glance a power decay like the Zipf law [27] $P\sim 1/K$ looks to be satisfactory. The formal power law fit $P\sim 1/K^{z},P^{*}\sim 1/(K^{*})^{z^{*}}$, done in log–log-scale for $1\leq K,K^{*}\leqslant q10$, gives the exponents $z=0.85\pm0.09,z^{*}=0.45\pm0.09$ (Fig. 7a), $z=0.88\pm0.10,z^{*}=0.77\pm0.16$ (Fig. 7b). However, the error bars for these fits are relatively large. Also other statistical tests (e.g. the Kolmogorov-Smirnov test, see details in [28]) give low statistical accuracy (e.g. statistical probability $p\approx0.2;0.1$ and $p\approx0.01;0.01$ for exponents $z,z^{*}=0.79,0.42$ and $0.75,0.65$ in Fig. 7a and Fig. 7b respectively). It is clear that 10 cultures is too small to have a good statistical accuracy. Thus, a larger number of cultures should be used to check the validity of the generalized Zipf law with a certain exponent. We make a conjecture that the Zipf law with the generalized exponents z,z^{*} will work in a better way for a larger number of multilingual Wikipedia editions which now have about 250 languages.

The distributions of cultures on the PageRank - CheiRank plane (K,K^{*}) are shown in Fig. 8. For the network of cultures constructed from top 30 PageRank persons we obtain the following ranking. The node WR is located at the top PageRank $K=1$ and it stays at the last CheiRank position $K^{*}=10$. This happens due to the fact that such persons as *Carl Linnaeus, Jesus, Aristotle, Plato, Alexander the Great, Muhammad* are not native for our 9 Wikipedia editions so that we have many nodes pointing to WR node, while WR has no outgoing links. The next node in PageRank is FR node at $K=2,K^{*}=5$, then DE node at $K=3,K^{*}=4$ and only then we find EN node at $K=4,K^{*}=7$. The node EN is not at all at top PageRank positions since it has many American politicians that does not count for links between cultures. After the world WR the top position is taken by French (FR) and then German (DE) cultures which have strong links inside the continental Europe.

However, the ranking is drastically changed when we consider top 30 2DRank persons. Here, the dominant role is played by art and science with singers, artists and scientists. The world WR here remains at the same position at $K=1, K^*=10$ but then we obtain English EN $(K=2, K^*=1)$ and German DE $(K=3, K^*=5)$ cultures while FR is moved to $K=K^*=7$.

Discussion

We investigated cross-cultural diversity of Wikipedia via ranking of Wikipedia articles. Even if the used ranking algorithms are purely based on network structure of Wikipedia articles, we find cultural distinctions and entanglement of cultures obtained from the multilingual editions of Wikipedia.

In particular, we analyze raking of articles about persons and identify activity field of persons and cultures to which persons belong. Politics is dominant in top PageRank persons, art is dominant in top 2DRank persons and in top CheiRank persons art and sport are dominant. We find that each Wikipedia edition favors its own persons, who have same cultural background, but there are also cross-cultural non-local heroes, and even ''global heroes''. We establish that local heroes are cited more often but non-local heroes on average are cited by more important articles.

Attributing top persons of the ranking list to different cultures we construct the network of cultures and characterize entanglement of cultures on the basis of Google matrix analysis of this directed network.

We considered only 9 Wikipedia editions selecting top 30 persons in a ''manual'' style. It would be useful to analyze a larger number of editions using an automatic computerized selection of persons from prefabricated listing in many languages developing lines discussed in [14]. This will allow to analyze a large number of persons improving the statistical accuracy of links between different cultures.

The importance of understanding of cultural diversity in globalized world is growing. Our computational, data driven approach can provide a quantitative and efficient way to understand diversity of cultures by using data created by millions of Wikipedia users. We believe that our results shed a new light on how organized interactions and links between different cultures.

Supporting Information

File S1 Presents Figures S1, S2, S3 in SI file showing comparison between probability distributions over activity fields and language for top 30 and 100 persons for EN, IT, NK respectively; tables S1, S2, … S27 in SI file showing top 30 persons in PageRank, CheiRank and 2DRank for all 9 Wikipedia editions. All names are given in English. Supplementary methods, tables, ranking lists and figures are available at http://www.quantware.ups-tlse.fr/QWLIB/wikiculturenetwork/; data sets of 9 hyperlink networks are available at [29] by a direct request addressed to S.Vigna.
(PDF)

Acknowledgments

We thank Sebastiano Vigna [29] who kindly provided to us the network data of 9 Wikipedia editions, collected in the frame of FET NADINE project. We thank Pablo Aragón and Andreas Kaltenbrunner for the list of persons in EN, IT, NL which we used to obtain supporting Figs.S1,S2,S3 in File S1.

Author Contributions

Conceived and designed the experiments: DLS. Performed the experiments: YHE. Analyzed the data: YHE DLS. Contributed reagents/materials/analysis tools: YHE DLS. Wrote the paper: YHE DLS.

References

1. Borges JL (1962) *The Library of Babel* in *Ficciones*, Grove Press, New York
2. Kaltenbrunner A, Laniado D (2012) *There is no deadline - time evolution of Wikipedia discussions*, Proc. of the 8th Intl. Symposium on Wikis and Open Collaboration, Wik- iSym12, Linz
3. Torok J, Iniguez G, Yasseri T, San Miguel M, Kaski K, et al. (2013) *Opinion, conflicts and consensus: modeling social dynamics in a collaborative enviroment* Phys Rev Lett 110: 088701
4. Yasseri T, Kornai A, Kertész J (2012) *A practical approach to language complexity: a Wikipedia case study* PLoS ONE, 7: e48386
5. Brandes U, Kenis P, Lerner U, van Raaij D (2009) *Network analysis of collaboration structure in Wikipedia* Proc. 18th Intl. Conf. WWW, :731
6. Hecht B, Gergle D (2009) *Measuring self-focus bias in community-maintained knowledge repositories* Proc. of the Fourth Intl Conf. Communities and technologies, ACM, New York :11
7. Nemoto K, Gloor PA (2011) *Analyzing cultural differences in collaborative innovation networks by analyzing editing behavior in different-language Wikipedias* Procedia - Social and Behavioral Sciences 26: 180
8. Norenzayan A (2011) *Explaining human behavioral diversity*, Science, 332: 1041
9. Gelfand MJ, Raver JL, Nishii L, Leslie LM, Lun J, et al. (2011) *Differences between tight and loose cultures: a 33-nation study*, Science, 332: 1100
10. Yasseri T, Spoerri A, Graham M, Kertész J (2013) *The most controversial topics in Wikipedia: a multilingual and geographical analysis* arXiv:1305.5566 [physics.soc-ph]
11. UNESCO World Report (2009) *Investing in cultural diversity and intercultural dialogue*, Available: http://www.unesco.org/new/en/culture/resources/report/the-unesco-world-report- on-cultural-diversity
12. Callahan ES, Herriing SC (2011) *Cultural bias in Wikipedia content on famous persons*, Journal of the American society for information science and technology 62: 1899
13. Warncke-Wang M, Uduwage A, Dong Z, Riedl J (2012) *In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network*, Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym 2012), ACM, New York No 20
14. Aragón P, Laniado D, Kaltenbrunner A, Volkovich Y (2012) *Biographical social networks on Wikipedia: a cross-cultural study of links that made history*, Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym 2012), ACM, New York No 19; arXiv:1204.3799v2[cs.SI]
15. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine Computer Networks and ISDN Systems 30: 107
16. Chepelianskii AD (2010) *Towards physical laws for software architecture* arXiv:1003.5455 [cs.SE]
17. Zhirov AO, Zhirov OV, Shepelyansky DL (2010) *Two-dimensional ranking of Wikipedia articles*, Eur Phys J B 77: 523
18. Ermann L, Chepelianskii AD, Shepelyansky DL (2012) *Toward two-dimensional search engines*, J Phys A: Math Theor 45: 275101
19. Eom YH, Frahm KM, Benczúr A, Shepelyansky DL (2013) *Time evolution of Wikipedia network ranking* arXiv:1304.6601 [physics.soc-ph]
20. Langville AM, Meyer CD (2006) *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton
21. Brin M, Stuck G (2002) *Introduction to dynamical systems*, Cambridge Univ. Press, Cambridge, UK
22. Ermann L, Frahm KM, Shepelyansky DL (2013) *Spectral properties of Google matrix of Wikipedia and other networks*, Eur Phys J D 86: 193
23. Chen P, Xie H, Maslov S, Redner S (2007) *Finding scientific gems with Googles PageRank algorithm* Jour Informetrics, 1: 8
24. Kwak H, Lee C, Park H, Moon S (2010) *What is Twitter, a social network or a news media?*, Proc. 19th Int. Conf. WWW2010, ACM, New York :591
25. Kandiah V, Shepelyansky DL (2013) *Google matrix analysis of DNA sequences*, PLoS ONE 8(5): e61519
26. Ermann L, Shepelyansky DL (2011) *Google matrix of the world trade network*, Acta Physica Polonica A 120(6A), A158
27. Zipf GK (1949) *Human behavior and the principle of least effort*, Addison-Wesley, Boston
28. Clauset A, Shalizi CR, Newman MEJ (2009) *Power-law distributions in empirical data*, SIAM Rev 51(4): 661
29. Personal website of Sebastiano Vigna. Available: http://vigna.dsi.unimi.it/. Accessed 2013 Jun 26.

SUPPORTING INFORMATION FOR: Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles

Young-Ho Eom[1], Dima L. Shepelyansky[1,*]

1 *Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France*

∗ Webpage: www.quantware.ups-tlse.fr/dima

1 Additional data

Supplementary methods, tables, ranking lists and figures are available at
http://www.quantware.ups-tlse.fr/QWLIB/wikiculturenetwork/;
data sets of 9 hyperlink networks are available at
http://vigna.dsi.unimi.it/
by a direct request addressed to S.Vigna.

Here we present additional figures and tables for the main part of the paper.

Figures S1, S2, S3 show comparison between probability distributions over activity fields and language for top 30 and 100 persons for EN, IT, NK respectively.

Tables show top 30 persons in PageRank, CheiRank and 2DRank for all 9 Wikipedia editions. All names are given in English.

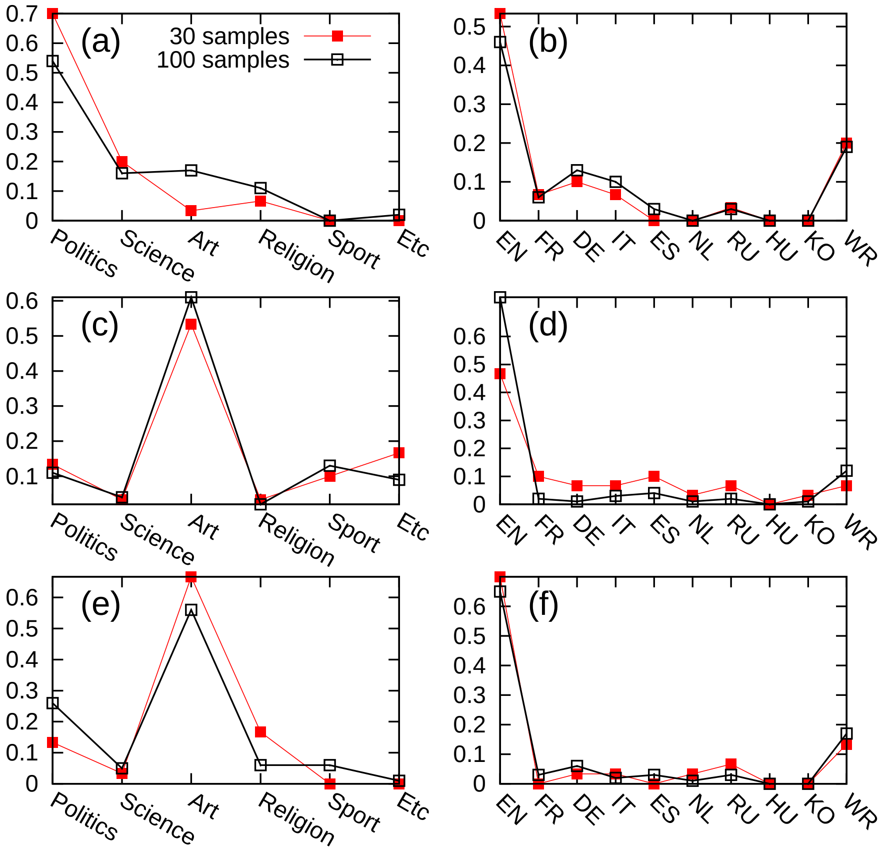


Figure S1: Probability distributions of activity fields and languages of top 30 persons and top 100 persons in English Wikipedia EN (total probability is normalized to unity): (a) Distribution of activity fields of PageRank top persons (b) Distribution of langauge of PageRank top persons. (c) Distribution of activity fields of CheiRank top persons (d) Distribution of langauge of CheiRank top persons. (e) Distribution of activity fields of 2DRank top persons (f) Distribution of langauge of 2DRank top persons.

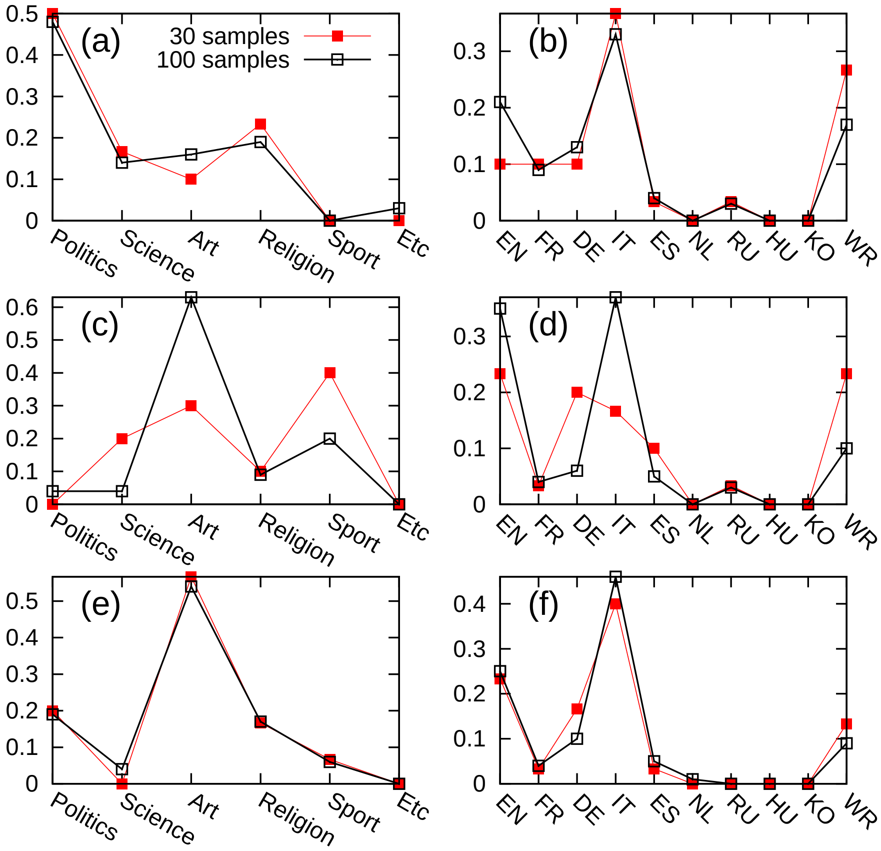


Figure S2: Same as in Fig.SI1 for Italian Wikipedia IT.

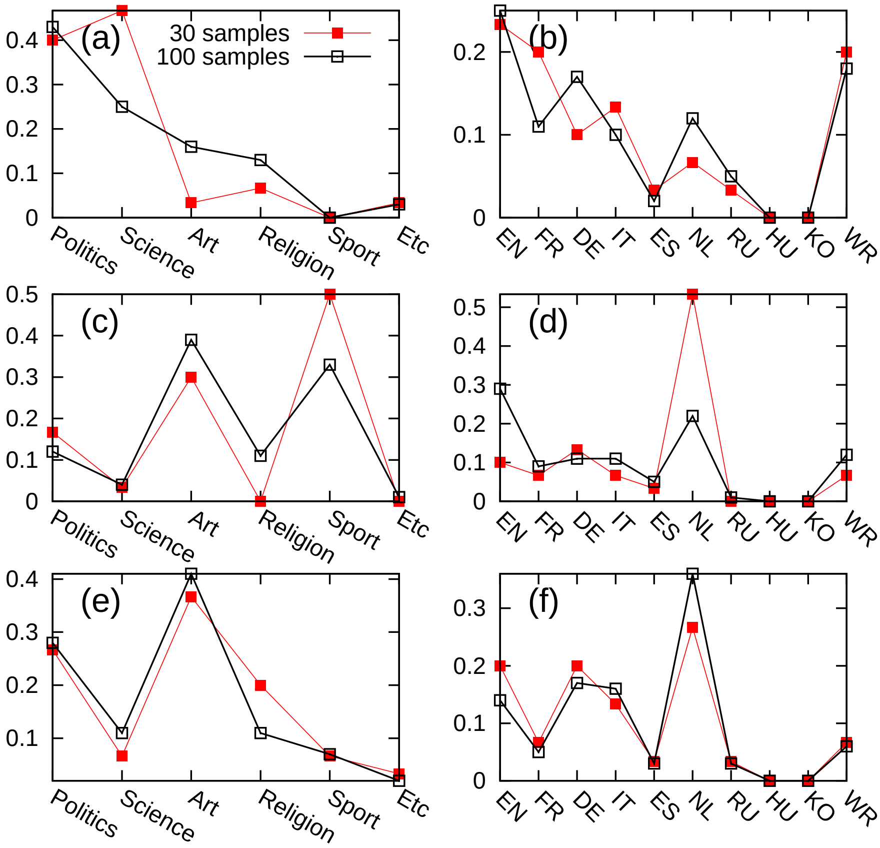


Figure S3: Same as in Fig.SI1 for Dutch Wikipedia NL.

Table S1: Top 30 persons by PageRank for English Wikipedia with their field of activity and native language.

$R_{EN,PageRank}$	Person	Field	Culture
1	Napoleon	Politics	FR
2	Carl Linnaeus	Science	WR
3	George W. Bush	Politics	EN
4	Barack Obama	Politics	EN
5	Elizabeth II	Politics	EN
6	Jesus	Religion	WR
7	William Shakespeare	Art	EN
8	Aristotle	Science	WR
9	Adolf Hitler	Politics	DE
10	Bill Clinton	Politics	EN
11	Franklin D. Roosevelt	Politics	EN
12	Ronald Reagan	Politics	EN
13	George Washington	Politics	EN
14	Plato	Science	WR
15	Richard Nixon	Politics	EN
16	Abraham Lincoln	Politics	EN
17	Joseph Stalin	Politics	RU
18	Winston Churchill	Politics	EN
19	John F. Kennedy	Politics	EN
20	Henry VIII of England	Politics	EN
21	Muhammad	Religion	WR
22	Thomas Jefferson	Politics	EN
23	Albert Einstein	Science	DE
24	Alexander the Great	Politics	WR
25	Augustus	Politics	IT
26	Charlemagne	Politics	FR
27	Karl Marx	Science	DE
28	Charles Darwin	Science	EN
29	Elizabeth I of England	Politics	EN
30	Julius Caesar	Politics	IT

Table S2: Top 30 persons by 2DRank for English Wikipedia with their field of activity and native language.

$R_{EN,2DRank}$	Person	Field	Culture
1	Frank Sinatra	Art	EN
2	Paul McCartney	Art	EN
3	Michael Jackson	Art	EN
4	Steven Spielberg	Art	EN
5	Pope Pius XII	Religion	IT
6	Vladimir Putin	Politics	RU
7	Mariah Carey	Art	EN
8	John Kerry	Politics	EN
9	Isaac Asimov	Art	EN
10	Stephen King	Art	EN
11	Dolly Parton	Art	EN
12	Prince (musician)	Art	EN
13	Robert Brown (botanist)	Science	EN
14	Vincent van Gogh	Art	NL
15	Lady Gaga	Art	EN
16	Beyoncé Knowles	Art	EN
17	Pope John Paul II	Religion	WR
18	Lord Byron	Art	EN
19	Muhammad	Religion	WR
20	Johnny Cash	Art	EN
21	Alice Cooper	Art	EN
22	Catherine the Great	Politics	RU
23	14th Dalai Lama	Religion	WR
24	Christina Aguilera	Art	EN
25	Marilyn Monroe	Art	EN
26	David Bowie	Art	EN
27	John McCain	Politics	EN
28	Bob Dylan	Art	EN
29	Johann Sebastian Bach	Art	DE
30	Jesus	Religion	WR

Table S2: Top 30 persons by CheiRank for English Wikipedia with their field of activity and native language.

$R_{EN,CheiRank}$	Person	Field	Culture
1	Roger Calmel	Art	FR
2	C. H. Vijayashankar	Politics	EN
3	Matt Kelley	ETC	EN
4	Alberto Cavallari	ETC	IT
5	Yury Chernavsky	Art	RU
6	William Shakespeare (inventor)	ETC	EN
7	Kelly Clarkson	Art	EN
8	Park Ji-Sung	Sport	KO
9	Mithun Chakraborty	Art	EN
10	Olga Sedakova	Sport	RU
11	Sara García	Art	ES
12	Pope Pius XII	Religion	IT
13	Andy Kerr	Politics	EN
14	Joe-Max Moore	Sport	EN
15	Josef Kemr	Art	WR
16	Darius Milhaud	Art	FR
17	Jan Crull, Jr.	ETC	EN
18	Farshad Fotouhi	Science	EN
19	Swaroop Kanchi	Art	EN
20	Jacques Lancelot	Art	FR
21	František Martin Pecháček	Art	DE
22	George Stephanekoulosech	ETC	EN
23	Chano Urueta	Art	ES
24	Franz Pecháček	Art	DE
25	Nicolae Iorga	Politics	WR
26	Arnold Houbraken	Art	NL
27	August Derleth	Art	EN
28	Javier Solana	Politics	ES
29	Drew Barrymore	Art	EN
30	Kevin Bloody Wilson	Art	EN

Table S4: Top 30 persons by PageRank for French Wikipedia with their field of activity and native language.

$R_{FR,PageRank}$	Person	Field	Culture
1	Napoleon	Politics	FR
2	Carl Linnaeus	Science	WR
3	Louis XIV of France	Politics	FR
4	Jesus	Religion	WR
5	Aristotle	Science	WR
6	Julius Caesar	Politics	IT
7	Charles de Gaulle	Politics	FR
8	Pope John Paul II	Religion	WR
9	Adolf Hitler	Politics	DE
10	Plato	Science	WR
11	Charlemagne	Politics	FR
12	Joseph Stalin	Politics	RU
13	Charles V, Holy Roman Emperor	Politics	ES
14	Napoleon III	Politics	FR
15	Nicolas Sarkozy	Politics	FR
16	Franois Mitterrand	Politics	FR
17	Victor Hugo	Art	FR
18	Jacques Chirac	Politics	FR
19	Honore de Balzac	Art	FR
20	Mary (mother of Jesus)	Religion	WR
21	Voltaire	Art	FR
22	George W. Bush	Politics	EN
23	Elizabeth II	Politics	EN
24	Muhammad	Religion	WR
25	Francis I of France	Politics	FR
26	William Shakespeare	Art	EN
27	Louis XVI of France	Politics	FR
28	Rene Descartes	Science	FR
29	Karl Marx	Science	DE
30	Louis XV of France	Politics	FR

Table S5: Top 30 persons by 2DRank for French Wikipedia with their field of activity and native language.

$R_{FR,2DRank}$	Person	Field	Culture
1	Franois Mitterrand	Politics	FR
2	Jacques Chirac	Politics	FR
3	Honore de Balzac	Art	FR
4	Nicolas Sarkozy	Politics	FR
5	Napoleon III	Politics	FR
6	Otto von Bismarck	Politics	DE
7	Michael Jackson	Art	EN
8	Adolf Hitler	Politics	DE
9	Ludwig van Beethoven	Art	DE
10	Johnny Hallyday	Art	FR
11	Napoleon	Politics	FR
12	Leonardo da Vinci	Art	IT
13	Jules Verne	Art	FR
14	Jacques-Louis David	Art	FR
15	Thomas Jefferson	Politics	EN
16	Sigmund Freud	Science	DE
17	Madonna (entertainer)	Art	EN
18	Serge Gainsbourg	Art	FR
19	14th Dalai Lama	Religion	WR
20	Alfred Hitchcock	Art	EN
21	Georges Clemenceau	Politics	FR
22	Carl Linnaeus	Science	WR
23	Steven Spielberg	Art	EN
24	J. R. R. Tolkien	Art	EN
25	Arthur Rimbaud	Art	FR
26	Charles Darwin	Science	EN
27	Maximilien de Robespierre	Politics	FR
28	Nelson Mandela	Politics	WR
29	Henry IV of France	Politics	FR
30	Charles de Gaulle	Politics	FR

Table S6: Top 30 persons by CheiRank for French Wikipedia with their field of activity and native language.

$R_{FR,CheiRank}$	Person	Field	Culture
1	John Douglas Lynch	Science	EN
2	Roger Federer	Sport	DE
3	Richard Upjohn Light	Science	EN
4	Jacques Davy Duperron	Art	FR
5	Rafael Nadal	Sport	ES
6	Martina Navratilova	Sport	EN
7	Michael Ilmari Saaristo	Science	WR
8	Kevin Bacon	Art	EN
9	Jean Baptiste Eble	Etc	FR
10	Marie-Magdeleine Ayme de La Chevreliere	Politics	FR
11	Nataliya Pyhyda	Sport	RU
12	Max Wolf	Science	DE
13	14th Dalai Lama	Religion	WR
14	Francoise Hardy	Art	FR
15	Ghislaine N. H. Sathoud	Etc	FR
16	Frank Glaw	Science	DE
17	Johnny Hallyday	Art	FR
18	Juan A. Rivero	Science	ES
19	Valentino Rossi	Sport	IT
20	Sheila (singer)	Art	FR
21	Franois Mitterrand	Politics	FR
22	Christopher Walken	Art	EN
23	Georges Clemenceau	Politics	FR
24	Elgin Loren Elwais	Sport	WR
25	Otto von Bismarck	Politics	DE
26	Edward Drinker Cope	Science	EN
27	Rashidi Yekini	Sport	WR
28	Tofiri Kibuuka	Sport	WR
29	Paola Espinosa	Sport	ES
30	Aksana Drahun	Sport	RU

Table S7: Top 30 persons by PageRank for German Wikipedia with their field of activity and native language.

$R_{DE,PageRank}$	Person	Field	Culture
1	Napoleon	Politics	FR
2	Carl Linnaeus	Science	WR
3	Adolf Hitler	Politics	DE
4	Aristotle	Science	WR
5	Johann Wolfgang von Goethe	Art	DE
6	Martin Luther	Religion	DE
7	Jesus	Religion	WR
8	Immanuel Kant	Science	DE
9	Charlemagne	Politics	FR
10	Plato	Science	WR
11	Pope John Paul II	Religion	WR
12	Karl Marx	Science	DE
13	Julius Caesar	Politics	IT
14	Augustus	Politics	IT
15	Louis XIV of France	Politics	FR
16	Friedrich Schiller	Art	DE
17	Wolfgang Amadeus Mozart	Art	DE
18	William Shakespeare	Art	EN
19	Josef Stalin	Politics	RU
20	Pope Benedict XVI	Religion	DE
21	Otto von Bismarck	Politics	DE
22	Cicero	Politics	IT
23	Wilhelm II, German Emperor	Politics	DE
24	Johann Sebastian Bach	Art	DE
25	Max Weber	Science	DE
26	Charles V, Holy Roman Emperor	Politics	ES
27	Frederick the Great	Politics	DE
28	Georg Wilhelm Friedrich Hegel	Science	DE
29	Mary (mother of Jesus)	Religion	WR
30	Augustine of Hippo	Religion	WR

Table S8: Top 30 persons by 2DRank for German Wikipedia with their field of activity and native language.

$R_{DE,2DRank}$	Person	Field	Culture
1	Adolf Hitler	Politics	DE
2	Otto von Bismarck	Politics	DE
3	Pope Paul VI	Religion	IT
4	Ludwig van Beethoven	Art	DE
5	Franz Kafka	Art	DE
6	George Frideric Handel	Art	DE
7	Gerhart Hauptmann	Art	DE
8	Bob Dylan	Art	EN
9	Johann Sebastian Bach	Art	DE
10	Alexander the Great	Politics	WR
11	Martin Luther	Religion	DE
12	Julius Caesar	Politics	IT
13	Joseph Beuys	Art	DE
14	Pope Leo XIII	Religion	IT
15	Carl Friedrich Gauss	Science	DE
16	Andy Warhol	Art	EN
17	Alfred Hitchcock	Art	EN
18	Thomas Mann	Art	DE
19	John Lennon	Art	EN
20	Augustus II the Strong	Politics	DE
21	Pope Benedict XVI	Religion	DE
22	Ferdinand II of Aragon	Politics	ES
23	Arthur Schnitzler	Art	DE
24	Martin Heidegger	Science	DE
25	Albrecht Dürer	Art	DE
26	Carl Linnaeus	Science	WR
27	Pablo Picasso	Art	ES
28	Rainer Werner Fassbinder	Art	DE
29	Wolfgang Amadeus Mozart	Art	DE
30	Historical Jesus	Religion	WR

Table S9: Top 30 persons by CheiRank for German Wikipedia with their field of activity and native language.

$R_{DE,CheiRank}$	Person	Field	Culture
1	Diomede Carafa	Religion	IT
2	Harry Pepl	Art	DE
3	Marc Zwiebler	Sport	DE
4	Eugen Richter	Politics	DE
5	John of Nepomuk	Religion	WR
6	Pope Marcellus II	Religion	IT
7	Karl Wilhelm Reinmuth	Science	WR
8	Johannes Molzahn	Art	DE
9	Georges Vanier	ETC	FR
10	Arthur Willibald Königsheim	ETC	DE
11	Thomas Fitzsimons	Politics	EN
12	Nelson W. Aldrich	Politics	EN
13	Ma Jun	ETC	WR
14	Michael Psellos	Religion	WR
15	Adolf Hitler	Politics	DE
16	Edoardo Fazzioli	ETC	IT
17	Ray Knepper	Sport	EN
18	Frédéric de Lafresnaye	Science	FR
19	Joan Crawford	Art	EN
20	Stephen King	Art	EN
21	Gerhart Hauptmann	Art	DE
22	Paul Moder	Politics	DE
23	Erni Mangold	Art	DE
24	Robert Stolz	Art	DE
25	Otto von Bismarck	Politics	DE
26	Christine Holstein	Art	DE
27	Pope Paul VI	Religion	IT
28	Franz Buxbaum	Science	DE
29	Gustaf Gründgens	Art	DE
30	Ludwig van Beethoven	Art	DE

Table S10: Top 30 persons by PageRank for Italian Wikipedia with their field of activity and native language.

$R_{IT,PageRank}$	Person	Field	Culture
1	Napoleon	Politics	FR
2	Jesus	Religion	WR
3	Aristotle	Science	WR
4	Augustus	Politics	IT
5	Pope John Paul II	Religion	WR
6	Dante Alighieri	Art	IT
7	Adolf Hitler	Politics	DE
8	Julius Caesar	Politics	IT
9	Benito Mussolini	Politics	IT
10	Charlemagne	Politics	FR
11	Mary (mother of Jesus)	Religion	WR
12	Plato	Science	WR
13	Isaac Newton	Science	EN
14	Charles V, Holy Roman Emperor	Politics	ES
15	Galileo Galilei	Science	IT
16	Louis XIV of France	Politics	FR
17	Constantine the Great	Politics	IT
18	Cicero	Politics	IT
19	Alexander the Great	Politics	WR
20	Paul the Apostle	Politics	WR
21	Albert Einstein	Science	DE
22	Joseph Stalin	Politics	RU
23	George W. Bush	Politics	EN
24	Silvio Berlusconi	Politics	IT
25	William Shakespeare	Art	EN
26	Augustine of Hippo	Religion	WR
27	Pope Paul VI	Religion	IT
28	Pope Benedict XVI	Religion	DE
29	Giuseppe Garibaldi	Politics	IT
30	Leonardo da Vinci	Science	IT

Table S11: Top 30 persons by 2DRank for Italian Wikipedia with their field of activity and native language.

$R_{IT,2DRank}$	Person	Field	Culture
1	Pope John Paul II	Religion	WR
2	Pope Benedict XVI	Religion	DE
3	Giuseppe Garibaldi	Politics	IT
4	Raphael	Art	IT
5	Jesus	Religion	WR
6	Benito Mussolini	Politics	IT
7	Michelangelo	Art	IT
8	Leonardo da Vinci	Art	IT
9	Pier Paolo Pasolini	Art	IT
10	Michael Jackson	Art	EN
11	Martina Navratilova	Sport	EN
12	Saint Peter	Religion	WR
13	Pope Paul III	Religion	IT
14	Wolfgang Amadeus Mozart	Art	DE
15	John Lennon	Art	EN
16	Bob Dylan	Art	EN
17	Mina (singer)	Art	IT
18	William Shakespeare	Art	EN
19	Julius Caesar	Politics	IT
20	Titian	Art	IT
21	Silvio Berlusconi	Politics	IT
22	Alexander the Great	Politics	WR
23	Pablo Picasso	Art	ES
24	Antonio Vivaldi	Art	IT
25	Ludwig van Beethoven	Art	DE
26	Napoleon	Politics	FR
27	Madonna (entertainer)	Art	EN
28	Roger Federer	Sport	DE
29	Johann Sebastian Bach	Art	DE
30	Walt Disney	Art	EN

Table S12: Top 30 persons by CheiRank for Italian Wikipedia with their field of activity and native language.

$R_{IT,CheiRank}$	Person	Field	Culture
1	Ticone di Amato	Religion	WR
2	John the Merciful	Religion	WR
3	Nduccio	Art	IT
4	Vincenzo Olivieri	Art	IT
5	Leo Baeck	Religion	DE
6	Karl Wilhelm Reinmuth	Science	DE
7	Freimut Börngen	Science	DE
8	Nikolai Chernykh	Science	RU
9	Edward L. G. Bowell	Science	EN
10	Roger Federer	Sport	DE
11	Michel Morganella	Sport	WR
12	Rafael Nadal	Sport	ES
13	Robin Söderling	Sport	WR
14	Iván Zamorano	Sport	ES
15	Martina Navratilova	Sport	EN
16	Venus Williams	Sport	EN
17	Goran Ivanišević	Sport	WR
18	Javier Pastore	Sport	ES
19	Stevan Jovetić	Sport	WR
20	Mina (singer)	Art	IT
21	George Ade	Art	EN
22	Kazuro Watanabe	Sport	WR
23	Andy Roddick	Sport	EN
24	Johann Strauss II	Art	DE
25	Max Wolf	Science	DE
26	Isaac Asimov	Art	EN
27	Georges Simenon	Art	FR
28	Alice Joyce	Art	EN
29	Pietro De Sensi	Sport	IT
30	Noemi (singer)	Art	IT

Table S13: Top 30 persons by PageRank for Spanish Wikipedia with their field of activity and native language.

$R_{ES,PageRank}$	Person	Field	Culture
1	Carl Linnaeus	Scinece	WR
2	Napoleon	Politics	FR
3	Jesus	Religion	WR
4	Aristotle	Science	WR
5	Charles V, Holy Roman Emperor	Politics	ES
6	Adolf Hitler	Politics	DE
7	Julius Caesar	Politics	IT
8	Philip II of Spain	Politics	ES
9	William Shakespeare	Art	EN
10	Plato	Science	WR
11	Albert Einstein	Science	DE
12	Augustus	Politics	IT
13	Pope John Paul II	Religion	WR
14	Christopher Columbus	ETC	IT
15	Karl Marx	Science	DE
16	Alexander the Great	Politics	WR
17	Isaac Newton	Science	EN
18	Francisco Franco	Politics	ES
19	Charlemagne	Politics	FR
20	Immanuel Kant	Science	DE
21	Charles Darwin	Science	EN
22	Louis XIV of France	Politics	FR
23	Mary (mother of Jesus)	Religion	WR
24	Wolfgang Amadeus Mozart	Art	DE
25	Galileo Galilei	Science	IT
26	Cicero	Politics	IT
27	Homer	Art	WR
28	Paul the Apostle	Religion	WR
29	René Descartes	Science	FR
30	Miguel de Cervantes	Art	ES

Table S14: Top 30 persons by 2DRank for Spanish Wikipedia with their field of activity and native language.

$R_{ES,2DRank}$	Person	Field	Culture
1	Wolfgang Amadeus Mozart	Art	DE
2	Julius Caesar	Politics	IT
3	Simón Bolívar	Politics	ES
4	Francisco Goya	Art	ES
5	Madonna (entertainer)	Art	EN
6	Bob Dylan	Art	EN
7	Barack Obama	Politics	EN
8	Fidel Castro	Politics	ES
9	Michael Jackson	Art	EN
10	Richard Wagner	Art	DE
11	Augusto Pinochet	Politics	ES
12	Trajan	Politics	IT
13	Jorge Luis Borges	Art	ES
14	Juan Perón	Politics	ES
15	Porfirio Díaz	Politics	ES
16	Michelangelo	Art	IT
17	J. R. R. Tolkien	Art	EN
18	Paul McCartney	Art	EN
19	Adolf Hitler	Politics	DE
20	John Lennon	Art	EN
21	Hugo Chávez	Politics	ES
22	Elizabeth II	Politics	EN
23	Lope de Vega	Art	ES
24	Francisco Franco	Politics	ES
25	Christopher Columbus	ETC	IT
26	Diego Velázquez	Art	ES
27	Pablo Picasso	Art	ES
28	Edgar Allan Poe	Art	EN
29	Charlemagne	Politics	FR
30	Juan Carlos I of Spain	Politics	ES

Table S15: Top 30 persons by CheiRank for Spanish Wikipedia with their field of activity and native language.

$R_{ES,CheiRank}$	Person	Field	Culture
1	Max Wolf	Science	DE
2	Monica Bellucci	Art	IT
3	Che Guevara	Politics	ES
4	Steve Buscemi	Art	EN
5	Johann Palisa	Science	DE
6	Auguste Charlois	Science	FR
7	José Flávio Pessoa de Barros	Science	WR
8	Arturo Mercado	Art	ES
9	Francisco Goya	Art	ES
10	Bob Dylan	Art	EN
11	Jorge Luis Borges	Art	ES
12	Brian May	Art	EN
13	Virgilio Barco Vargas	Politics	ES
14	Mariano Bellver	ETC	ES
15	Demi Lovato	Art	EN
16	Joan Manuel Serrat	Art	ES
17	Mary Shelley	Art	EN
18	Ana Belén	Art	ES
19	Aki Misato	Art	WR
20	Carl Jung	Science	DE
21	Roger Federer	Sport	DE
22	Antoni Gaudí	Art	ES
23	Rafael Nadal	Sport	ES
24	Hans Melchior	Science	DE
25	Paulina Rubio	Art	ES
26	Paul McCartney	Art	EN
27	Julieta Venegas	Art	ES
28	Fermin Muguruza	Art	ES
29	Belinda (entertainer)	Art	ES
30	Patricia Acevedo	Art	ES

Table S16: Top 30 persons by PageRank for Dutch Wikipedia with their field of activity and native language.

$R_{NL,PageRank}$	Person	Field	Culture
1	Carl Linnaeus	Science	WR
2	Pierre Andre Latreille	Science	FR
3	Napoleon	Politics	FR
4	Eugene Simon	Science	FR
5	Jesus	Religion	WR
6	Charles Darwin	Science	EN
7	Julius Caesar	Politics	IT
8	Adolf Hitler	Politics	DE
9	Aristotle	Science	WR
10	Charlemagne	Politics	FR
11	Plato	Science	WR
12	Jean-Baptiste Lamarck	Science	FR
13	Ernst Mayr	Science	DE
14	Alexander the Great	Politics	WR
15	Louis XIV of France	Politics	FR
16	Pope John Paul II	Religion	WR
17	Alfred Russel Wallace	Science	EN
18	Charles V, Holy Roman Emperor	Politics	ES
19	Thomas Robert Malthus	Science	EN
20	Augustus	Politics	IT
21	William I of the Netherlands	Politics	NL
22	Joseph Stalin	Politics	RU
23	Albert Einstein	Science	DE
24	Beatrix of the Netherlands	Politics	NL
25	Christopher Columbus	Etc	IT
26	Elizabeth II	Politics	EN
27	Isaac Newton	Science	EN
28	Wolfgang Amadeus Mozart	Art	DE
29	J. B. S. Haldane	Science	EN
30	Cicero	Politics	IT

Table S17: Top 30 persons by 2DRank for Dutch Wikipedia with their field of activity and native language.

$R_{NL,2DRank}$	Person	Field	Culture
1	Pope Benedict XVI	Religion	DE
2	Elizabeth II	Politics	EN
3	Charles Darwin	Science	EN
4	Albert II of Belgium	Politics	NL
5	Albert Einstein	Science	DE
6	Pope John Paul II	Religion	WR
7	Michael Jackson	Art	EN
8	Johann Sebastian Bach	Art	DE
9	Saint Peter	Religion	WR
10	Johan Cruyff	Sport	NL
11	William Shakespeare	Art	EN
12	Christopher Columbus	Etc	IT
13	Augustus	Politics	IT
14	Frederick the Great	Politics	DE
15	Rembrandt	Art	NL
16	Eddy Merckx	Sport	NL
17	Ludwig van Beethoven	Art	DE
18	Pope Pius XII	Religion	IT
19	Peter Paul Rubens	Art	NL
20	Napoleon	Politics	FR
21	Wolfgang Amadeus Mozart	Art	DE
22	Igor Stravinsky	Art	RU
23	Martin of Tours	Religion	FR
24	Geert Wilders	Politics	NL
25	J.R.R. Tolkien	Art	EN
26	Pierre Cuypers	Art	NL
27	Charles V, Holy Roman Emperor	Politics	ES
28	Pope Pius IX	Religion	IT
29	Juliana of the Netherlands	Politics	NL
30	Elvis Presley	Art	EN

Table S18: Top 30 persons by CheiRank for Dutch Wikipedia with their field of activity and native language.

$R_{NL,CheiRank}$	Person	Field	Culture
1	Pier Luigi Bersani	Politics	IT
2	Francesco Rutelli	Politics	IT
3	Hans Renders	Science	NL
4	Julian Jenner	Sport	NL
5	Marten Toonder	Art	NL
6	Uwe Seeler	Sport	DE
7	Stefanie Sun	Art	WR
8	Roger Federer	Sport	DE
9	Theo Janssen	Sport	NL
10	Zazie	Art	FR
11	Albert II of Belgium	Politics	NL
12	Denny Landzaat	Sport	NL
13	Paul Biegel	Art	NL
14	Guido De Padt	Politics	NL
15	Jan Knippenberg	Sport	NL
16	Michael Schumacher	Sport	DE
17	Hans Werner Henze	Art	DE
18	Lionel Messi	Sport	ES
19	Johan Cruijff	Sport	NL
20	Eva Janssen (actrice)	Art	NL
21	Marion Zimmer Bradley	Art	EN
22	Graham Hill	Sport	EN
23	Rick Wakeman	Art	EN
24	Mihai Nesu	Sport	NL
25	Freddy De Chou	Politics	NL
26	Rubens Barrichello	Sport	WR
27	Ismail Aissati	Sport	NL
28	Marco van Basten	Sport	NL
29	Paul Geerts	Art	NL
30	Ibrahim Afellay	Sport	NL

Table S19: Top 30 persons by PageRank for Russian Wikipedia with their field of activity and native language.

$R_{RU,PageRank}$	Person	Field	Culture
1	Peter the Great	Politics	RU
2	Napoleon	Politics	FR
3	Carl Linnaeus	Science	WR
4	Joseph Stalin	Politics	RU
5	Alexander Pushkin	Art	RU
6	Vladimir Lenin	Politics	RU
7	Catherine the Great	Politics	RU
8	Jesus	Religion	WR
9	Aristotle	Science	WR
10	Vladimir Putin	Politics	RU
11	Julius Caesar	Politics	IT
12	Adolf Hitler	Politics	DE
13	Boris Yeltsin	Politics	RU
14	William Shakespeare	Art	EN
15	Ivan the Terrible	Politics	RU
16	Alexander II of Russia	Politics	RU
17	Nicholas II of Russia	Politics	RU
18	Karl Marx	Science	DE
19	Louis XIV of France	Politics	FR
20	Nicholas I of Russia	Politics	RU
21	Alexander I of Russia	Politics	RU
22	Alexander the Great	Politics	WR
23	Charlemagne	Politics	FR
24	William Herschel	Science	EN
25	Mikhail Gorbachev	Politics	RU
26	Paul I of Russia	Politics	RU
27	Leo Tolstoy	Art	RU
28	Nikolai Gogol	Art	RU
29	Dmitry Medvedev	Politics	RU
30	Lomonosov	Science	RU

Table S20: Top 30 persons by 2DRank for Russian Wikipedia with their field of activity and native language.

$R_{RU,2DRank}$	Person	Field	Culture
1	Dmitri Mendeleev	Science	RU
2	Peter the Great	Politics	RU
3	Justinian I	Politics	WR
4	Yaroslav the Wise	Politics	RU
5	Elvis Presley	Art	EN
6	Yuri Gagarin	Etc	RU
7	William Shakespeare	Art	EN
8	Albert Einstein	Science	DE
9	Adolf Hitler	Politics	DE
10	Christopher Columbus	Etc	IT
11	Catherine the Great	Politics	RU
12	Vladimir Vysotsky	Art	RU
13	Louis de Funes	Art	FR
14	Lomonosov	Science	RU
15	Alla Pugacheva	Art	RU
16	Viktor Yanukovych	Politics	RU
17	Nikolai Gogol	Art	RU
18	Felix Dzerzhinsky	Politics	RU
19	Aleksandr Solzhenitsyn	Art	RU
20	Pope Benedict XVI	Religion	DE
21	Maxim Gorky	Art	RU
22	Julius Caesar	Politics	IT
23	George Harrison	Art	EN
24	Bohdan Khmelnytsky	Politics	RU
25	Rembrandt	Art	NL
26	John Lennon	Art	EN
27	Jules Verne	Art	FR
28	Benito Mussolini	Politics	IT
29	Nicholas Roerich	Art	RU
30	Niels Bohr	Science	WR

Table S21: Top 30 persons by CheiRank for Russian Wikipedia with their field of activity and native language.

$R_{RU,CheiRank}$	Person	Field	Culture
1	Aleksander Vladimirovich Sotnik	Etc	RU
2	Aleksei Aleksandrovich Bobrinsky	Politics	RU
3	Boris Grebenshchikov	Art	RU
4	Karl Wilhelm Reinmuth	Science	DE
5	Ronnie O'Sullivan	Sport	EN
6	Max Wol	Science	DE
7	Ivan Egorovich Sizykh	Etc	RU
8	Vladimir Mikhilovich Popkov	Art	RU
9	Sun Myung Moon	Religion	KO
10	Mikhail Pavlovich Tolstoi	Etc	RU
11	Perry Como	Art	EN
12	John Heenan	Religion	EN
13	Petr Aleksandrovich Ivaschenko	Art	RU
14	Andrey Vlasov	Etc	RU
15	Christian Heinrich Friedrich Peters	Science	DE
16	Auguste Charlois	Science	FR
17	Damian (Marczhuk)	Religion	RU
18	Yuri Gagarin	Etc	RU
19	Stephen Hendry	Sport	EN
20	Ivan Grigorevich Donskikh	Etc	RU
21	Anna Semenovna Kamenkova-Pavlova	Art	RU
22	Ivan Nikolaevich Shulga	Art	RU
23	George Dwyer	Religion	EN
24	William Wheeler (bishop)	Religion	EN
25	Vladimir Vladimirovitsch Antonik	Art	RU
26	Leonid Parfyonov	Art	RU
27	Vincent Nichols	Religion	EN
28	Dmitri Mendeleev	Science	RU
29	Boris Vladimirovich Bakin	Etc	RU
30	George Harrison	Art	EN

Table S22: Top 30 persons by PageRank for Hungarian Wikipedia with their field of activity and native language.

$R_{HU,PageRank}$	Person	Field	Culture
1	Carl Linnaeus	Science	WR
2	Jesus	Religion	WR
3	Napoleon	Politics	FR
4	Aristotle	Science	WR
5	Julius Caesar	Politics	IT
6	Matthias Corvinus	Politics	HU
7	Szentagothai Janos	Science	HU
8	William Shakespeare	Art	EN
9	Adolf Hitler	Politics	DE
10	Stephen I of Hungary	Politics	HU
11	Augustus	Politics	IT
12	Michael Schumacher	Sport	DE
13	Miklos Rethelyi	Politics	HU
14	Sigismund, Holy Roman Emperor	Politics	HU
15	Lajos Kossuth	Politics	HU
16	Charles I of Hungary	Politics	HU
17	Bela IV of Hungary	Politics	HU
18	Maria Theresa	Politics	DE
19	Joseph Stalin	Politics	RU
20	Franz Joseph I of Austria	Politics	DE
21	Louis I of Hungary	Politics	HU
22	Francis II Rakoczi	Politics	HU
23	Mary (mother of Jesus)	Religion	WR
24	Sandor Petofi	Art	HU
25	Pope John Paul II	Religion	WR
26	Johann Wolfgang von Goethe	Art	DE
27	Alexander the Great	Politics	WR
28	Bela Bartok	Art	HU
29	Charlemagne	Politics	FR
30	Louis XIV of France	Politics	FR

Table S23: Top 30 persons by 2DRank for Hungarian Wikipedia with their field of activity and native language.

$R_{HU,2DRank}$	Person	Field	Culture
1	Stephen I of Hungary	Politics	HU
2	Sandor Petofi	Art	HU
3	Franz Liszt	Art	HU
4	Kati Kovacs	Art	HU
5	Alexander the Great	Politics	WR
6	Attila Jozsef	Art	HU
7	Aristotle	Science	WR
8	Kimi Raikkonen	Sport	WR
9	Rubens Barrichello	Sport	WR
10	Lajos Kossuth	Politics	HU
11	Bela Bartok	Art	HU
12	Charlemagne	Politics	FR
13	Sandor Weores	Art	HU
14	Mariah Carey	Art	EN
15	Wolfgang Amadeus Mozart	Art	DE
16	Josip Broz Tito	Politics	WR
17	Charles I of Hungary	Politics	HU
18	Isaac Asimov	Art	EN
19	Napoleon	Politics	FR
20	Bonnie Tyler	Art	EN
21	Miklos Radnoti	Art	HU
22	Jay Chou	Art	WR
23	Janos Kodolanyi	Art	HU
24	Louis I of Hungary	Politics	HU
25	Zsuzsa Koncz	Art	HU
26	Adolf Hitler	Politics	HU
27	Stephen King	Art	EN
28	Mor Jokai	Art	HU
29	Ferenc Erkel	Art	HU
30	Franz Joseph I of Austria	Politics	DE

Table S24: Top 30 persons by CheiRank for Hungarian Wikipedia with their field of activity and native language.

$R_{HU,CheiRank}$	Person	Field	Culture
1	Edward L. G. Bowell	Science	EN
2	Karl Wilhelm Reinmuth	Science	DE
3	Max Wolf	Science	DE
4	Benjamin Boukpeti	Sport	FR
5	Urata Takesi	Science	WR
6	Wilfred Bungei	Sport	WR
7	Henri Debehogne	Science	FR
8	Lee "Scratch" Perry	Art	WR
9	Karl Golsdorf	Etc	DE
10	Johann Palisa	Science	DE
11	Dirk Kuijt	Sport	NL
12	Roger Federer	Sport	DE
13	Csernus Imre	Etc	HU
14	Kati Kovacs	Art	HU
15	Rafael Nadal	Sport	ES
16	Venus Williams	Sport	EN
17	Sebastien Loeb	Sport	FR
18	Pleh Csaba	Science	HU
19	Tibor Antalpeter	Sport	HU
20	Serena Williams	Sport	EN
21	Csore Gabor	Art	HU
22	Pirmin Schwegler	Sport	DE
23	Olivia Newton-John	Art	EN
24	Petter Solberg	Sport	WR
25	Orosz Anna	Art	HU
26	Zsambeki Gabor	Art	HU
27	Vera Igorevna Zvonarjova	Sport	RU
28	Sandor Petofi	Art	HU
29	Roberta Vinci	Sport	IT
30	Flavia Pennetta	Sport	HU

Table S25: Top 30 persons by PageRank for Korean Wikipedia with their field of activity and native language.

$R_{KO,PageRank}$	Person	Field	Culture
1	Carl Linnaeus	Science	WR
2	Gojong of the Korean Empire	Politics	KO
3	Jesus	Religion	WR
4	John Edward Gray	Science	EN
5	Aristotle	Science	WR
6	Napoleon	Politics	FR
7	Sejong the Great	Politics	KO
8	Park Chung-hee	Politics	KO
9	Emperor Wu of Han	Politics	WR
10	Seonjo of Joseon	Politics	KO
11	Taejong of Joseon	Politics	KO
12	Syngman Rhee	Politics	KO
13	Kim Dae-jung	Politics	KO
14	Roh Moo-hyun	Politics	KO
15	Yeongjo of Joseon	Politics	KO
16	Adolf Hitler	Politics	DE
17	Taejo of Joseon	Politics	KO
18	Sukjong of Joseon	Politics	KO
19	Kim Il-sung	Politics	KO
20	Qianlong Emperor	Politics	WR
21	Kim Jong-il	Politics	KO
22	Kangxi Emperor	Politics	WR
23	Emperor Gaozu of Han	Politics	WR
24	Chun Doo-hwan	Politics	KO
25	Taejo of Goryeo	Politics	KO
26	George W. Bush	Politics	EN
27	Qin Shi Huang	Politics	WR
28	Jeongjo of Joseon	Politics	KO
29	Sunjo of Joseon	Politics	KO
30	Cao Cao	Politics	WR

Table S26: Top 30 persons by 2DRank for Korean Wikipedia with their field of activity and native language.

$R_{KO,2DRank}$	Person	Field	Culture
1	Gojong of the Korean Empire	Politics	KO
2	Sejong the Great	Politics	KO
3	Park Chung-hee	Politics	KO
4	Taejong of Joseon	Politics	KO
5	Kim Dae-jung	Politics	KO
6	Roh Moo-hyun	Politics	KO
7	Syngman Rhee	Politics	KO
8	Kim Il-sung	Politics	KO
9	Qianlong Emperor	Politics	WR
10	Kangxi Emperor	Politics	WR
11	Taejo of Goryeo	Politics	KO
12	Seonjo of Joseon	Politics	KO
13	Jeongjo of Joseon	Politics	KO
14	Kim Young-sam	Politics	KO
15	Julius Caesar	Politics	IT
16	Chun Doo-hwan	Politics	KO
17	Injo of Joseon	Politics	KO
18	Tokugawa Ieyasu	Politics	WR
19	Lee Myung-bak	Politics	KO
20	Seongjong of Joseon	Politics	KO
21	Cao Cao	Politics	WR
22	Confucius	Science	WR
23	Mao Zedong	Politics	WR
24	Taejo of Joseon	Politics	KO
25	Toyotomi Hideyoshi	Politics	WR
26	Heungseon Daewongun	Politics	KO
27	Liu Bei	Politics	WR
28	Yeongjo of Joseon	Politics	KO
29	Pope John Paul II	Religion	WR
30	Adolf Hitler	Politics	DE

Table S27: Top 30 persons by CheiRank for Korean Wikipedia with their field of activity and native language.

$R_{KO,CheiRank}$	Person	Field	Culture
1	Lee Jong-wook (baseball)	Sport	KO
2	Kim Dae-jung	Politics	KO
3	Lionel Messi	Sport	ES
4	Kim Kyu-sik	Politics	KO
5	Johannes Kepler	Science	DE
6	Yun Chi-young	Politics	KO
7	Michael Jackson	Art	EN
8	Yi Sun-sin	ETC	KO
9	Chang Myon	Politics	KO
10	IU (singer)	Art	KO
11	Kim Seo-yeong	Art	KO
12	Tokugawa Ieyasu	Politics	WR
13	Jeremy Renner	Art	EN
14	Zhao Deyin	Politics	WR
15	Yang Joon-Hyu	Sport	KO
16	Zhang Gui (Tang Dynasty)	Politics	WR
17	Zinedine Zidane	Sport	FR
18	Park Chung-hee	Politics	KO
19	Heungseon Daewongun	Politics	KO
20	Ahn Ji-hwan	Art	KO
21	Lee Seung-Yeop	Sport	KO
22	Roh Moo-hyun	Politics	KO
23	Britney Spears	Art	EN
24	Kim Young-sam	Politics	KO
25	Jeong Hyeong-don	Art	KO
26	Kim Yu-Na	Sport	KO
27	Park Jong-Seol	Art	KO
28	Lim Taekyoung	Art	KO
29	Park Ji-Sung	Sport	KO
30	Yuh Woon-Hyung	Politics	KO

Cross-Lingual Web Spam Classification

András Garzó Bálint Daróczy Tamás Kiss Dávid Siklósi András A. Benczúr

Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)
Eötvös University, Budapest
{garzo, daroczyb, kisstom, sdavid, benczur}@ilab.sztaki.hu

ABSTRACT

While Web spam training data exists in English, we face an expensive human labeling procedure if we want to filter a Web domain in a different language. In this paper we overview how existing content and link based classification techniques work, how models can be "translated" from English into another language, and how language-dependent and independent methods combine. In particular we show that simple bag-of-words translation works very well and in this procedure we may also rely on mixed language Web hosts, i.e. those that contain an English translation of part of the local language text. Our experiments are conducted on the ClueWeb09 corpus as the training English collection and a large Portuguese crawl of the Portuguese Web Archive. To foster further research, we provide labels and precomputed values of term frequencies, content and link based features for both ClueWeb09 and the Portuguese data.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; I.2 [**Computing Methodologies**]: Artificial Intelligence

General Terms

Document Classification, Information Retrieval, Hyperlink Analysis

Keywords

Cross-lingual text processing, Web classification. Web spam, Content analysis, Link analysis

1. INTRODUCTION

It has already been known from the early results on text classification that "obtaining classification labels is expensive" [32]. This is especially true in multilingual collections where either separate training labels have to be produced for each language in question, or techniques of cross-lingual information retrieval [13] or machine translation [35] have to be used.

While several results focus on cross-lingual classification of general text corpora [2; 38; 43, and many more], we concentrate on the special and characteristically different problem of Web classification. Web spam filtering, the area of devising methods to identify useless Web content with the sole purpose of manipulating search engine results, has drawn much attention in the past years [41, 29, 26]. Our results on cross-lingual Web classification are motivated by the needs and opportunities of Internet archives [4].

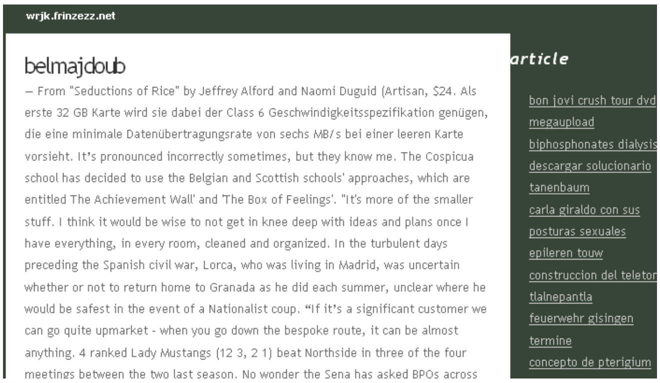
Web classification may exploit methods of recent evaluation campaigns on Web spam filtering. Our results combine methods from two areas, cross-lingual information retrieval and Web classification. Traditional methods in cross-lingual information retrieval use dictionaries, machine translation methods, and more recently multilingual Wikipedia editions. Web classification on the other hand relies on features of content and linkage [9], some of which are language independent. However, language independence does not necessarily imply domain independence: PageRank and its variants may have different distributions for differing interconnectivity and the ratio of the "boundary": the pages not included but pointed to by some page in the domain, crawl, or language. TrustRank and query popularity based features depend on the availability of a trusted seed set, typically hosts listed in the Open Directory Project (`http://dmoz.org`), and the coverage of search queries. Finally, the typical word length and text entropy may also vary language by language.

This paper experiments with a new combination of learning methods and cross-lingual features for web classification. Our task is different from standard methods of cross-lingual text classification (see [43] and references therein) in the following aspects:

- We classify hosts not individual pages as this is the standard task for Web spam [7].
- Even if we consider a national domain, the actual language used in a host can be mixed, especially for spam pages automatically generated from chunks (see Fig. 1 as an example).
- We may exploit multilingualism by classifying a host based on its part written in English.

We note that host level classification is preferred for Web spam filtering due to the facts that (1) fine-grained page or even comment level classification is computationally unfeasible on the Web scale; and (2) the goal is to filter mass amounts of spam including link farms and machine generated content that can be blocked on the host level. Indeed, our set of labeled Portuguese spam hosts is the byproduct of the normal quality assessment procedure conducted within the Portuguese Web Archive. In previous results [9; 7, and many more] full host names are used as a domain and we use this definition in this paper, however we argue that a

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.



wrjk.frinzezz.net

belmajdoub

— From "Seductions of Rice" by Jeffrey Alford and Naomi Duguid (Artisan, $24. Als erste 32 GB Karte wird sie dabei der Class 6 Geschwindigkeitsspezifikation genügen, die eine minimale Datenübertragungsrate von sechs MB/s bei einer leeren Karte vorsieht. It's pronounced incorrectly sometimes, but they know me. The Cospicua school has decided to use the Belgian and Scottish schools' approaches, which are entitled The Achievement Wall' and 'The Box of Feelings'. "It's more of the smaller stuff. I think it would be wise to not get in knee deep with ideas and plans once I have everything, in every room, cleaned and organized. In the turbulent days preceding the Spanish civil war, Lorca, who was living in Madrid, was uncertain whether or not to return home to Granada as he did each summer, unclear where he would be safest in the event of a Nationalist coup. "If it's a significant customer we can go quite upmarket - when you go down the bespoke route, it can be almost anything. 4 ranked Lady Mustangs (12 3, 2 1) beat Northside in three of the four meetings between the two last season. No wonder the Sena has asked BPOs across

article

- bon jovi crush tour dvd
- megaupload
- biphosphonates dialysis
- descargar solucionario
- tanenbaum
- carla giraldo con sus
- posturas sexuales
- epileren touw
- construccion del teleton
- tlalnepantla
- feuerwehr gisingen
- termine
- concepto de pterigium

Figure 1: Portion of a mixed language machine generated spam page.

pay level domain or even IP based definition [15] would fit the problem even better. In addition, labeling a page or an entire host is almost the same effort for a human, and very frequently a single page cannot even be assessed without seeing the context, very much unlike email spam or spam in social media.

In this paper we investigate how much various classes of Web content and linkage features, some requiring very high computational effort, add to the classification accuracy. As the bag of words representation turned out to describe Web hosts best for most classification tasks of the ECML/PKDD 2010 Discovery Challenge [15], we realized that new text classification methods are needed for the cross-lingual task.

Based on recent results in Web spam filtering, we also collect and handle a large number of features and test a variety of machine learning techniques, including SVM, ensemble selection, LogitBoost and Random Forest. Our key findings are summarized next.

- Hosts that contain a mix of English and national language content, likely translations, yield a very powerful resource for cross-lingual classification. Some of our methods work even without using dictionaries, not to mention without more complex tools of natural language processing.
- Similar to our previous English-only results, the bag-of-words representation together with appropriate machine learning techniques is the strongest method for Web classification.
- The "public" spam features of Castillo et al. [9], especially the content-based ones, depend heavily on the data collection and have little generalizational power. For spam classification they require cross-corpus normalization while for topical classification, the content based features do not seem to be applicable.

To assess the prediction power of the proposed features, we run experiments over the `.pt` domain [24, 23]. Our techniques are evaluated along several alternatives and yield a considerable improvement in terms of area-under-the-curve (AUC).

The rest of the paper is organized as follows. After a review of related research at the intersection of machine learning, cross-lingual information retrieval and Web mining (Section 2), we introduce the proposed learning methods and describe the classification features (Section 3). Our experimental results are presented in Section 4.

2. RELATED WORK

We base our methods on results of both cross-lingual and Web classification that we review next. In general, cross-lingual classification either works by translating documents [38, 30, 43], or terms only [2], or using an intermediate language-independent representation of concepts [44]. For general results on cross-lingual text classification we refer to [2] who propose linguistic resources such as dictionaries similar to the ones used in cross-lingual information retrieval. As a broad overview, we refer to the CLEF Ad Hoc tasks overview papers, e.g. [13] in the latter area. We also note that several results exploit Wikipedia linkage and local editions [42, 31, 22].

Several cross-language classification results, similar to ours, work over "pseudo-English" documents by translating key terms into English using dictionaries [2], or using latent semantic analysis [14, 36]. The cross-lingual classification results reported are however, unlike ours, much worse than the monolingual baselines.

Semi-supervised learning finds applications in cross-lingual classification where, similar to our methods, the unlabeled part of the data is also used for building the model. Expectation maximization is used in [38, 39] for cleansing the classifier model from translation errors; others [37] exploit document similarities over the unlabeled corpus. In [43] co-training over machine translated Chinese and English text is used for sentiment analysis.

Closest to our goals is the method of [30] for classifying Chinese Web pages using English training data, however, either because of the cultural differences between Chinese and English content or the fact that they classified on the page and not host level, they achieve accuracy metrics much weaker than for the monolingual counterpart. We also note that they are aware of the existence of multilingual content but they apparently do not exploit the full power of multilingual hosts. Finally, a recent Web page classification method described in [44] uses matrix tri-factorization for learning an auxiliary language, an approach that we find computationally unfeasible for classification in the scale of a top level domain.

Text classification is studied extensively in classical information retrieval. While traditional term-based topical classification for Web content relies on local page content only, several solutions tailored to the web use terms from linked pages as well [5, 21]. Semi-supervised learning methods (surveyed, for instance, in [47]) exploit information from both labeled and unlabeled data instances. Relational learning methods (presented, for instance, in [20]) also consider existing relationships between data instances.

Recognizing and preventing spam has been identified as one of the top challenges for web search engines [29, 41]. As all major search engines use page, anchor text, and link analysis algorithms to produce their rankings of search results, web spam appears in sophisticated forms that manipulate page contents as well as the interconnection structure [27]. Accordingly, spam hunters also rely on a variety of content [17, 34, 18] and link [28, 3, 46] based features to detect web spam; a recent evaluation of their combination is provided in [9]. In the area of the so-called Adversarial Information Retrieval, workshop series ran for five years [16], evaluation campaigns including the Web Spam Challenges [7] were organized. The ECML/PKDD Discovery Challenge 2010 (see e.g. [15]) extended the scope by introducing labels for genre and quality by serving the needs of a fictional archive.

Our baseline classification procedures are collected by analyzing the results of the Web Spam Challenges and the ECML/PKDD Discovery Challenge 2010. Best results either used bag of words vectors or the so-called "public" feature sets of [8]. The Discovery Challenge 2010 best results [25, 1, 33] and our analysis [15, 40] show that the bag of words representation variants proved to be very strong for the English collection. For classification techniques, a wide selection including decision trees, random forest, SVM, class-feature-centroid, boosting, bagging and oversampling in addition to feature selection (Fisher, Wilcoxon, Information Gain) were used. In our previous work [40], we improved over the best results of the Challenge participants by the combination of SVM and biclustering over the bag of words representation of the hosts. These experiments indicate little use of link and content based features. A possible reason is that the DC2010 training and test sets were constructed in a way that no IP and domain was allowed to be split between training and testing. The rationale is that once a domain or IP is found to consist of spam, its subdomains or other hosts on the same server are much more likely spam and their classification becomes straightforward. This simple consideration was not implemented in early datasets: the Web Spam Challenge data sets were labeled by uniform random sampling. For this reason, we have to reconsider the applicability of propagation [46] and graph stacking [9].

3. METHOD

Our Web host classification applies a classifier ensemble consisting of features based on content and linkage as well as various English, translated, and semi-supervised Portuguese bag of words models. The following subsections describe the core ingredients. The standard content and link-based features[1] and the necessary transformations from the English to the Portuguese collection are described in Sections 3.1 and 3.2, respectively. In Section 3.3 we describe our bag-of-words translation method and SVM based classifiers, followed by a semi-supervised algorithm that relies on multilingual host content to first give prediction using a pure English model and then apply the results to train a Portuguese model. Finally the ensemble method ingredients are found in Section 3.5.

3.1 Features: Content

Among the early content spam papers, Fetterly et al. [17] demonstrated that a sizable portion of machine generated spam pages can be identified through statistical analysis. Ntoulas et al. [34] introduce a number of content based spam features including number of words in the page, title, anchor, as well as the fraction of page drawn from popular words, and the fraction of most popular words that appear in the page. Spam hunters use a variety of additional content based features [6, 18] to detect web spam; a recent measurement of their combination appears in [9] who also provide these methods as a *public feature set* for the Web Spam Challenges.

We use the public feature set [9] that includes the following values computed for the home page of the domain, the page with the maximum PageRank, and the average over the entire host:

1. Number of words in the page, title;
2. Average word length, average word trigram likelihood;
3. Compression rate, entropy;
4. Fraction of anchor text, visible text;
5. Corpus and query precision and recall.

Here feature classes 1–4 can be normalized by using the average and standard deviation values over the two collections, while class 4 is likely domain and language independent.

Corpus precision and recall are defined over the k most frequent words in the dataset, excluding stopwords. Corpus precision is the fraction of words in a page that appear in the set of popular terms while corpus recall is the fraction of popular terms that appear in the page. This class of features is language independent but rely on different lists of most frequent terms for the two data sets.

Query precision and recall is based on frequencies from query logs that have to be either compiled separately for each language or domain (questions from Portugal likely have different distribution than from Brazil), or the English query list has to be translated. Since we had no access to a query log in Portuguese, we selected the second approach.

3.2 Features: Linkage

Recently several results have appeared that apply rank propagation to extend initial judgments over a small set of seed pages or sites to the entire web, such as trust [28, 46] or distrust. These ideas were distilled into the public link based feature set [9] and include the following values with averages, standard deviation, and several functions computed from them:

- Assortativity, reciprocity;
- In and out-degree;
- Host and page neighborhood size at various distances;
- PageRank and truncated variants.

One of the strongest features is TrustRank [28], PageRank personalized on known honest hosts. TrustRank however needs a trusted seed set. Typically hosts that appear in the Open Directory Project (ODP) are used as seed. Unfortunately, ODP acts as our negative sample set as well, hence in this paper we have to omit TrustRank, one of the strongest link-based features in our discussion.

3.3 Features: Bag-of-Words

Spam can be classified purely based on the terms used. Based on our recent result, we use libSVM [10] with several kernels and apply late fusion as described in [40]. The bag of words representation of a Web host consists of the top 10,000 most frequent terms after stop word removal.

In order to classify hosts in Portuguese, we translate the Portuguese terms to construct an English bag of words representation of the host. The procedure is described in Algorithm 1 with the following considerations:

- Short terms are not translated as they typically cause noise and often coincide between the languages.

[1] `http://barcelona.research.yahoo.net/webspam/datasets/uk2007/features/`

Algorithm 1 Algorithm for translating Portuguese term counts for evaluation by an English model

```
for all Top 10,000 most frequent English terms $en$ do
    count[$en$] = count of term $en$ in host $h$
for all Top 10,000 most frequent Portuguese terms $pt$ of at least four letters do
    count_pt[$pt$] = count of term $pt$ in host $h$
    variants = number of single-term English translations of $pt$
    if variants > 0 then
        for all $en$: single-term Portuguese translations of $pt$ do
            count[$en$]+ = count_pt[$pt$]/variants
Classify $h$ using term counts count[$en$]
```

- Multiple translation alternatives exist. We consider all translations, but we split the term frequency value between them in order not to overweight terms with many translations. A smarter but more complex weighting method is described in [39].
- Multi-word translation, such as *Monday* through *Friday* translated into *Segunda* through *Sexta feira*, cannot be handled based on single term frequencies. Since counting expressions (multi-word sequences) would complicate the process, we omitted this step in our experiments.
- Portuguese terms may coincide with English ones and counted in the first **for** loop. And they may have no translation, in which case the term is omitted.

We use the BM25 term weighting scheme. Let there be H hosts consisting of an average $\overline{\ell}$ terms. Given a term t of frequency f over a given host that contains ℓ terms, the weight of t in the host becomes

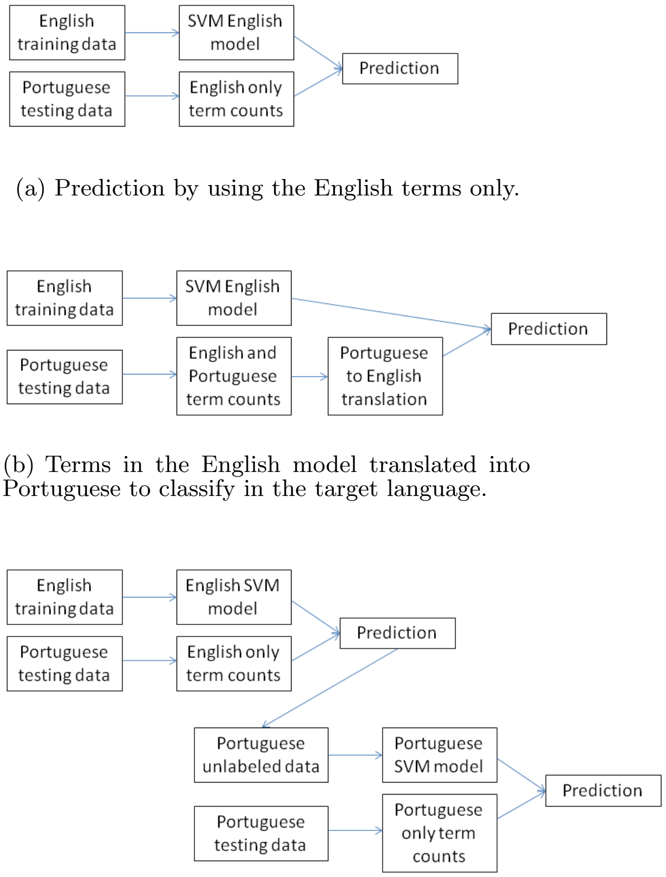
$$\log \frac{H - h + 0.5}{h + 0.5} \cdot \frac{f(k+1)}{f + k(1 - b + b \cdot \frac{\ell}{\overline{\ell}})}. \qquad (1)$$

This expression turned out to perform best in our earlier results [15]. As optimal parameters, an exceptionally low value $k = 1$ and a large $b = 0.5$ turned out to perform best in preliminary experiments. Low k means very quick saturation of the term frequency function while large b downweights content from very large Web hosts. We do not show extensive experiments on these parameters.

3.4 Semi-supervised cross-lingual learning based on multilingual Web sites

A large portion of national language Web content appears on the same host in English version as well, as seen in Fig. 3. This figure shows the proportion of the total frequency of the 10,000 most frequent Portuguese terms within the sum of the Portuguese and English top 10,000 frequencies. This fact gives rise to several options of English, Portuguese and mixed language text classification. As summarized in Fig. 2, the simplest solution is to ignore non-English content and simply use term frequencies of the most frequent English terms as measured over the English part of ClueWeb09. Another solution, as described in Section 3.3, is to translate the whole content term by term into English and use the model trained over ClueWeb09 again.

We may however rely on mixed language hosts to classify without using a dictionary in a semi-supervised proce-



(a) Prediction by using the English terms only.

(b) Terms in the English model translated into Portuguese to classify in the target language.

(c) After applying the method of Fig. 2(a), strongest positive and negative predictions are used for training a model in the target language.

Figure 2: Three methods for classifying mixed language content based on a monolingual training set.

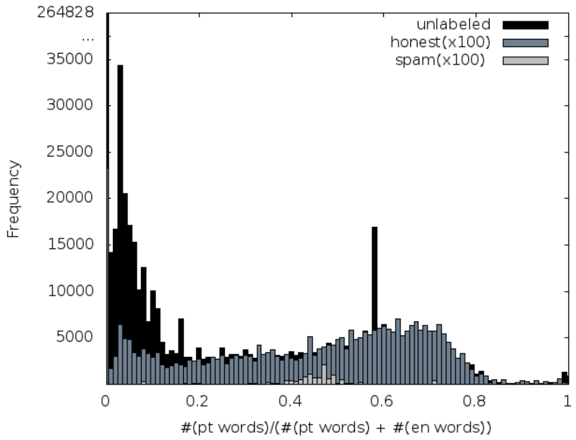


Figure 3: Statistics for the language distribution of most frequent terms in Web hosts over the .pt domain, with the 257,000 English-only hosts removed, separate for spam, ODP and unlabeled hosts. A very large fraction of the unlabeled hosts is English only, shown with a break in the horizontal scale.

dure using these (unlabeled) hosts. In Algorithm 2 we give a two-step stacked classification procedure summarized in Fig. 2(c). First we select hosts that contain an appropriate mix of English and Portuguese terms, the middle range in Fig. 3 between threshold_low = 0.4 and threshold_high = 0.6. Based on the English term frequencies of these hosts, we give prediction using a model trained over the English part of ClueWeb09. Now we turn to Portuguese term count based modeling. Even in the case when no labeled Portuguese training data exists, we may now use the outcome of the English model as training labels. More precisely, if a host has predicted value less than pred_low = 0.1, then we use the host as a negative, and if more than pred_high = 0.9, then as a positive training instance.

Algorithm 2 Stacked classification of mixed-language hosts based on an English model

```
for all hosts h do
  ratio[h] = total frequency of top 10,000 Portuguese
  terms divided by total frequency of top 10,000 Portuguese and English terms
  if threshold_low < ratio[h] < threshold_high then
    pred[h] = prediction for h based on the English model
    if pred[h] < pred_low then
      Add h to negative training instances
    if pred[h] > pred_high then
      Add h to positive training instances
Train a model based on Portuguese term counts using the
positive and negative instances h
Classify all testing h using the Portuguese only model
```

We select the intermediate training set efficiently by first running a MapReduce job only to count the dictionary term distribution, and then compute features for the selected hosts but not for the others.

We also note that the procedure summarized by the scheme in Fig. 2(c) can be used with any classifier and feature set. In addition to training using Portuguese term frequencies, we also compute the public content based features and compare models trained on ClueWeb09 vs. the semi-supervised "training" set.

3.5 Classification Framework

In our classifier ensemble we split features into related sets as described in Sections 3.1–3.3 and for each set we use a collection of classifiers that fit the data type and scale. These classifiers are then combined by ensemble selection. We used the classifier implementations of the machine learning toolkit Weka [45]. We use a procedure similar to [15] that we summarize here.

In the context of combining classifiers for Web classification, to our best knowledge, ensemble selection was only used by our previous result [15]. Before that, only simple methods that combine the predictions of SVM or decision tree classifiers through logistic regression or random forest have been used [11]. We believe that the ability to combine a large number of classifiers while preventing overfitting makes ensemble selection an ideal candidate for Web classification, since it allows us to use a large number of features and learn different aspects of the training data at the same time. Instead of tuning various parameters of different classifiers, we can concentrate on finding powerful features and selecting the main classifier models which we believe to be able to capture the differences between the classes to be distinguished.

We used Weka ensemble selection [45] for performing the experiments. We allow Weka to use all available models in the library for greedy sort initialization and use 5-fold embedded cross-validation during ensemble training and building. We set AUC as the target metric to optimize for and run 100 iterations of the hillclimbing algorithm.

We use the following model types for building the model library for ensemble selection: bagged and boosted decision trees, logistic regression, LogitBoost, naive Bayes and random forests. For most classes of features we use all classifiers and allow selection to choose the best ones. The exception is static term vector based features, where, due to the very large number of features, we use SVM as described in Sections 3.3–3.4.

4. EXPERIMENTS

We evaluate the performance of the proposed classification approach on a 2009 crawl of the Portuguese Web Archive of more than 600,000 domains and 70M pages. For training our English language models, we used the English part of ClueWeb09 of approximately 20M domains and 500M pages. Web spam labels were provided by the Portuguese Web Archive and the Waterloo Spam Rankings [12], respectively. While the Waterloo Spam Rankings contain negative training instances as well, for the Portuguese data we used pages from the Open Directory Project (ODP) for this purpose. The distribution of labels and the number of pages in labeled and all hosts is seen in Fig. 4. In our results we use the ClueWeb09 labels for training and the Portuguese Web Archive data for testing only, thus measuring the case when training only over English language labeled data.

We use the area under the ROC curve (AUC) [19] as used at Web Spam Challenge 2008 [7] to evaluate our classifiers. We do not give results in terms of precision, recall, F-measure or any other measure that depends on the selection of a threshold, as these measures are sensitive to the threshold and do not give a stable comparison of two results. These measures, to our best knowledge, were not used in Web classification evaluation campaigns since after Web Spam Challenge 2007.

4.1 Feature distributions

As seen by the language distribution in Fig. 3, our Portuguese testing data set consists of hosts with English to Portuguese ratio uniformly spread between mostly English to fully Portuguese, with the exception of a large number of English only hosts. These latter hosts are, however, underrepresented in the labeled set that we use for testing our cross-lingual method, hence we take no specific action to classify them.

Since most often web sites are topically classified based on the strong signals derived from terms that appear on their pages, our first and often most powerful classifier is SVM over tf.idf, averaged over all pages of the host. After stop word removal, we use the most frequent 10,000 terms both in English and in Portuguese.

The distribution of content features differs significantly between ClueWeb09 and the Portuguese crawl. As an example, the relative behavior of spam compared to normal hosts also significantly differs between ClueWeb09 and the

category	.pt count	ClueWeb count
spam	124	439
honest	3375	8421
hosts	686443	19228332
pages	71656081	502368557

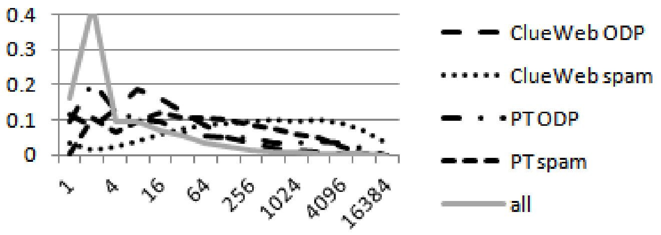


Figure 4: The number of positive and negative labeled host instances and the host and page count for the two data sets. The labeled ClueWeb data is identical to that of [12]. The chart on the right shows the fraction of labeled and all hosts with a given number of pages, with an exponential binning.

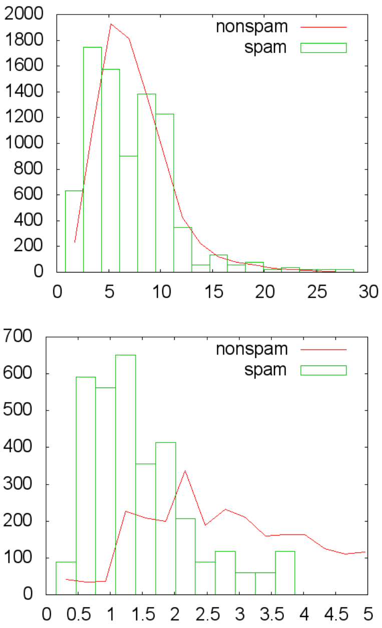


Figure 5: Distribution of the title length of the home page over the ClueWeb09 (top) and the Portuguese data (bottom), separate for spam and normal hosts.

Content ensemble	0.719
Content LogitBoost	0.751
Link	0.921
English	0.752
Translated	0.861
Stacked	0.894
Translated+Stacked avg	0.895
English+Stacked avg	0.899
English+Translated avg	0.952
English+Translated+Stacked avg	0.952
English+Link avg	0.898
Translated+Link avg	0.950
Translated+Stacked+Link avg	0.953
Stacked+Link avg	0.964
English+Translated+Link avg	0.967
English+Stacked+Link avg	0.976
English+Translated+Stacked+Link	0.976

Table 1: AUC of the main classification methods over the Portuguese test data. In the two variants of the content based features, we give results of the ensemble selection in the first and a single LogitBoost in the second column.

Portuguese data as seen in Fig. 5. Hence we may not expect content based features to work well across models.

4.2 Results

We show our results in terms of the AUC measure over the Portuguese Web test data set trained over the ClueWeb09 labels in Table 1. First, we give results obtained by using the public content and link based features [9]. These features work relative well for spam. Improved results are obtained by using LogitBoost only instead of the full classifier ensemble, as seen by comparing the first and second rows of Table 1. Link features (row 3) perform surprisingly well despite of the lack of TrustRank features.

The relative power of content and link based features over the training corpus is apparently similar. In our crossvalidation experiment over ClueWeb09, the training set, we obtain an AUC of 0.806 for content and 0.804 for linkage. For the Portuguese data, the link features trained over the ClueWeb09 corpus perform much better (0.921) than cross-validated over the training data itself. This may be due to the fact that labeled spam comes from a relative small number of link farms and hence have a very characteristic link structure.

Next, we give our results based on the bag of words representation for training in English and using labels of the Portuguese collection only for testing. Considering the Portuguese corpus as it was written in English (row “English”) is clearly a bad idea, still its performance matches that of the content features. The translation model (row “Translated”) works much better than the fully English one and is further improved by the stacked framework of Section 3.4 (row “Stacked”). Finally, we combine subsets of the classifiers by averaging their predicted spamicity values. The first block contains all four combinations of the three bag of words methods (English, Translated and Stacked); and the second block in addition combines with the LogitBoost classifier output over the link features. The combination of all models except the Translated one is the overall best method (last two rows). Here we observe that the combination of the English and translated classifiers can only be beaten by using the linkage features. On the other hand the Stacked model combines very well with linkage and the final best result consists of their combination with the English classifier.

Conclusion

In the paper we have demonstrated the applicability of cross-lingual Web host level spam filtering. Our experiments were tested over more than 600,000 hosts of the `.pt` domain by using the near 20M host English part of the ClueWeb09 data sets. Our results open the possibility for Web classification practice in national Internet archives who are mainly concerned about their resources, require fast reacting methods, and have very limited budget for human assessment.

By our experiments it has turned out that the strongest resources for cross-lingual classification are linkage as well as multilingual Web sites that discuss the same topic in both English and the local language. Note that these Web sites cannot be considered parallel corpora: we have no guarantee of exact translations, however, as our experiments also indicate, their content in different languages are topically identical. The use of dictionaries to translate a bag of words based model also works and combine well with other methods. The normalization of the "public" Web spam content based features [9] across languages however seems to fail; also these features perform weak for topical classification. Link based features can however be used for language-independent Web spam classification, regardless of their weakness identified in our previous result [15].

To foster further research, we provide labels and precomputed values of term frequencies, content and link based features for both ClueWeb09 and the Portuguese data available at `http://datamining.sztaki.hu/en/crosslingual/`.

Acknowledgments

To Daniel Gomes and colleagues from the Portuguese Web Archive `http://arquivo.pt/` for providing us with the Portuguese labeled data set. To Julien Masanès, Philippe Rigaux and colleagues from the Internet Memory Foundation for drawing our attention to the relevance of the problem and testing preliminary versions of our method on their data.

Work conducted at the Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI) was supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956), by the EU FP7 Project LAWA—Longitudinal Analytics of Web Archive Data, and by the grant OTKA NK 105645.

The work of Tamás Kiss reported in the paper has been developed in the framework of the project "Talent care and cultivation in the scientific workshops of BME" project. This project is supported by the grant TÁMOP - 4.2.2.B-10/1–2010-0009.

Work conducted at the Eötvös University, Budapest was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013). The research was carried out as part of the EITKIC_12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group. (`www.ictlabs.elte.hu`)

5. REFERENCES

[1] L. D. Artem Sokolov, Tanguy Urvoy and O. Ricard. Madspam consortium at the ECML/PKDD discovery challenge 2010. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[2] N. Bel, C. Koster, and M. Villegas. Cross-lingual text categorization. *Research and Advanced Technology for Digital Libraries*, pages 126–139, 2003.

[3] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. SpamRank – Fully automatic link spam detection. In *Proc. 1st AIRWeb, held in conjunction with WWW2005*, 2005.

[4] A. A. Benczúr, M. Erdélyi, J. Masanés, and D. Siklósi. Web spam challenge proposal for filtering in archives. In *Proc. 5th AIRWeb, held in conjunction with WWW2009*. ACM Press, 2009.

[5] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[6] A. Bratko, B. Filipič, G. Cormack, T. Lynam, and B. Zupan. Spam Filtering Using Statistical Data Compression Models. *The Journal of Machine Learning Research*, 7:2673–2698, 2006.

[7] C. Castillo, K. Chellapilla, and L. Denoyer. Web spam challenge 2008. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[8] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

[9] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. *Proc. 30th ACM SIGIR*, pages 423–430, 2007.

[10] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[11] G. Cormack. Content-based Web Spam Detection. In *Proc. 3rd AIRWeb*, 2007.

[12] G. Cormack, M. Smucker, and C. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465, 2011.

[13] G. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. Clef 2007: Ad hoc track overview. *Advances in Multilingual and Multimodal Information Retrieval*, pages 13–32, 2008.

[14] S. Dumais, T. Letsche, M. Littman, and T. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21, 1997.

[15] M. Erdélyi, A. Garzó, and A. A. Benczúr. Web spam classification: a few features worth more. In *Joint WICOW/AIRWeb Workshop on Web Quality, in conjunction with WWW2011, Hyderabad, India.* ACM Press, 2011.

[16] D. Fetterly and Z. Gyöngyi. Fifth international workshop on adversarial information retrieval on the web (AIRWeb 2009). 2009.

[17] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics – Using statistical analysis to locate spam web pages. In *Proceedings of the 7th*

International Workshop on the Web and Databases (WebDB), pages 1–6, Paris, France, 2004.

[18] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In *Proc 28th ACM SIGIR*, Salvador, Brazil, 2005.

[19] J. Fogarty, R. S. Baker, and S. E. Hudson. Case studies in the use of roc curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005*, GI '05, pages 129–136, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2005. Canadian Human-Computer Communications Society.

[20] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[21] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using Web structure for classifying and describing Web pages. In *Proc. 11th WWW*, 2002.

[22] J. Göbölös-Szabó, N. Prytkova, M. Spaniol, and G. Weikum. Cross-lingual data quality for knowledge base acceleration across wikipedia editions. In *Proc. QDB*, 2012.

[23] D. Gomes, J. Miranda, and M. Costa. A survey on web archiving initiatives. In S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, editors, *Research and Advanced Technology for Digital Libraries*, LNCS vol. 6966, pages 408–420. Springer Berlin Heidelberg, 2011.

[24] D. Gomes, A. Nogueira, J. Miranda, and M. Costa. Introducing the portuguese web archive initiative. 2009.

[25] X.-C. Z. Guang-Gang Geng, Xiao-Bo Jin and D. Zhang. Evaluating web content quality via multi-scale features. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[26] Z. Gyöngyi and H. Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, October 2005.

[27] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc 1st AIRWeb*, Chiba, Japan, 2005.

[28] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. 30th VLDB*, pages 576–587, Toronto, Canada, 2004.

[29] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[30] X. Ling, G. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu. Can chinese web pages be classified with english data source? In *Proc. 17th WWW*, pages 969–978. ACM, 2008.

[31] X. Ni, J. Sun, J. Hu, and Z. Chen. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proc. fourth ACM WSDM*, pages 375–384. ACM, 2011.

[32] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134, 2000.

[33] V. Nikulin. Web-mining with wilcoxon-based feature selection, ensembling and multiple binary classifiers. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[34] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. 15th WWW*, pages 83–92, Edinburgh, Scotland, 2006.

[35] J. Olive, C. Christianson, and J. McCary. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation.* Springer, 2011.

[36] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proc. 48th ACL*, pages 1118–1127. Association for Computational Linguistics, 2010.

[37] G. Ramírez-de-la Rosa, M. Montes-y Gómez, L. Villasenor-Pineda, D. Pinto-Avendano, and T. Solorio. Using information from the target language to improve crosslingual text classification. *Advances in Natural Language Processing*, pages 305–313, 2010.

[38] L. Rigutini, M. Maggini, and B. Liu. An em based training algorithm for cross-language text categorization. In *Proc. 2005 IEEE/WIC/ACM Web Intelligence*, pages 529–535. IEEE, 2005.

[39] L. Shi, R. Mihalcea, and M. Tian. Cross language text classification by model translation and semi-supervised learning. In *Proc. EMNLP 2010*, pages 1057–1067. Association for Computational Linguistics, 2010.

[40] D. Siklósi, B. Daróczy, and A. Benczúr. Content-based trust and bias classification via biclustering. In *Proc. 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 41–47. ACM, 2012.

[41] A. Singhal. Challenges in running a commercial search engine. In *IBM Search and Collaboration Seminar 2004*. IBM Haifa Labs, 2004.

[42] P. Sorg and P. Cimiano. Enriching the crosslingual link structure of wikipedia-a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, pages 49–54, 2008.

[43] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics, 2009.

[44] H. Wang, H. Huang, F. Nie, and C. Ding. Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization. In *Proc. 34th ACM SIGIR*, pages 933–942. ACM, 2011.

[45] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.

[46] B. Wu, V. Goel, and B. D. Davison. Topical TrustRank: Using topicality to combat web spam. In *Proc. 15th WWW*, Edinburgh, Scotland, 2006.

[47] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

The classification power of Web features*

Miklós Erdélyi[1,2], András A. Benczúr[1,3], Bálint Daróczy[1,3], András Garzó[1,3], Tamás Kiss[1,3] and Dávid Siklósi[1,3]

[1]Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA SZTAKI)
[2]University of Pannonia, Department of Computer Science and Systems Technology, Veszprém
[3]Eötvös University Budapest

Abstract

In this paper we give a comprehensive overview of features devised for Web spam detection and investigate how much various classes, some requiring very high computational effort, add to the classification accuracy.

- We collect and handle a large number of features based on recent advances in Web spam filtering, including temporal ones, in particular we analyze the strength and sensitivity of linkage change.
- We propose new temporal link similarity based features and show how to compute them efficiently on large graphs.
- We show that machine learning techniques including ensemble selection, LogitBoost and Random Forest significantly improve accuracy.
- We conclude that, with appropriate learning techniques, a simple and computationally inexpensive feature subset outperforms all previous results published so far on our data set and can only slightly be further improved by computationally expensive features.
- We test our method on three major publicly available data sets, the Web Spam Challenge 2008 data set WEBSPAM-UK2007, the ECML/PKDD Discovery Challenge data set DC2010 and the Waterloo Spam Rankings for ClueWeb09.

*This work was supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956), by the EU FP7 Project LAWA—Longitudinal Analytics of Web Archive Data, OTKA NK 105645 and by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013). The research was carried out as part of the EITKIC_12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group. (www.ictlabs.elte.hu) This paper is a comprehensive comparison of the best performing classification techniques based on [9, 37, 36, 38] and new experiments.

Our classifier ensemble sets the strongest classification benchmark as compared to participants of the Web Spam and ECML/PKDD Discovery Challenges as well as the TREC Web track.

To foster research in the area, we make several feature sets and source codes public[1], including the temporal features of eight `.uk` crawl snapshots that include WEBSPAM-UK2007 as well as the Web Spam Challenge features for the labeled part of ClueWeb09.

1 Introduction

Web classification finds several use, both for content filtering and for building focused corpora from a large scale Web crawl. As one notable use, Internet archives actively participate in large scale experiments [8], some of them building analytics services over their collections [6]. Most of the existing results on Web classification originate from the area of Web spam filtering that have turned out to generalize to a wide class of tasks including genre, Open Directory category, as well as quality classification. Closely related areas include filtering and tagging in social networks [50].

Web spam filtering, the area of devising methods to identify useless Web content with the sole purpose of manipulating search engine results, has drawn much attention in the past years [63, 49, 46]. The first mention of Web spam, termed *spamdexing* as a combination of words *spam* and (search engine) *indexing*, appears probably in a 1996 news article [27] as part of the early Web era discussions on the spreading porn content [24]. In the area of the so-called Adversarial Information Retrieval workshop series ran since 2005 [40] and evaluation campaigns including the Web Spam Challenges [18], the ECML/PKDD Discovery Challenge 2010 [50] and the Spam task of TREC 2010 Web Track [29] were organized. A recent comprehensive survey on Web spam filtering research is found in [19].

In this paper we present, to our best knowledge, the most comprehensive experimentation based on content, link as well as temporal features, both new and recently published. Our spam filtering baseline classification procedures are collected by analyzing the results [28, 1, 44] of the Web Spam Challenges and the ECML/PKDD Discovery Challenge 2010 [45, 2, 58]. Our comparison is based on AUC values [42] that we believe to be more stable as it does not depend on the split point; indeed, while Web Spam Challenge 2007 used F-measure and AUC, Web Spam Challenge 2008 used AUC only as evaluation measure.

Web spam appears in sophisticated forms that manipulate content as well as linkage [47] with examples such as

- Copied content, “honey pots” that draw attention but link to unrelated, spam targets;
- Garbage content, stuffed with popular or monetizable query terms and phrases such as university degrees, online casinos, bad credit status or

[1] `https://datamining.sztaki.hu/en/download/web-spam-resources`

adult content;

- Link farms, a large number of strongly interlinked pages across several domains.

The Web spammer toolkit consists of a clearly identifiable set of manipulation techniques that has not changed much recently. The Web Spam Taxonomy of Gyöngyi et al. [47] distinguishes content (term) and link spamming along with techniques of hiding, cloaking and removing traces by e.g. obfuscated redirection. Most of the features designed fight either link or content spamming.

We realize that recent results have ignored the importance of the machine learning techniques and concentrated only on the definition of new features. Also the only earlier attempt to unify a large set of features [20] is already four years old and even there little comparison is given on the relative power of the feature sets. For classification techniques, a wide selection including decision trees, random forest, SVM, class-feature-centroid, boosting, bagging and oversampling in addition to feature selection (Fisher, Wilcoxon, Information Gain) were used [45, 2, 58] but never compared and combined. In this paper we address the following questions.

- Do we get the maximum value out of the features we have? Are we sufficiently sophisticated at applying machine learning?
- Is it worth calculating computationally expensive features, in particular some related to page-level linkage?
- What is an optimal feature set for a fast spam filter that can quickly react at crawl time after fetching a small sample of a Web site?

We compare our result with the very strong baselines of the Web Spam Challenge 2008 and ECML/PKDD 2010 Discovery Challenge data sets. Our main results are as follows.

- We apply state-of-the-art classification techniques by the lessons learned from KDD Cup 2009 [57]. Key in our performance is ensemble classification applied both over different feature subsets as well as over different classifiers over the same features. We also apply classifiers yet unexplored against Web spam, including Random Forest [14] and LogitBoost [43].
- We compile a small yet very efficient feature set that can be computed by sample pages from the site while completely ignoring linkage information. By this feature set a filter may quickly react to a recently discovered site and intercept in time before the crawler would start to follow a large number of pages from a link farm. This feature set itself reaches AUC 0.893 over WEBSPAM-UK2007.
- Last but not least we gain strong improvements over the Web Spam Challenge best performance [18]. Our best result in terms of AUC reaches 0.9 and improves on the best Discovery Challenge 2010 results.

Several recent papers propose temporal features [61, 55, 31, 52] to improve classification accuracy. We extend link-based similarity algorithms by proposing metrics to capture the linkage change of Web pages over time. We describe a method to calculate these metrics efficiently on the Web graph and then measure their performance when used as features in Web spam classification. We propose an extension of two link-based similarity measures: XJaccard and PSimRank [41].

We investigate the combination of temporal and non-temporal, both link- and content-based features using ensemble selection. We evaluate the performance of ensembles built on the latter feature sets and compare our results to that of state-of-the-art techniques reported on our dataset. Our conclusion is that temporal and link-based features in general do not significantly increase Web spam filtering accuracy. However, information about linkage change might improve the performance of a language independent classifier: the best results for the French and German classification tasks of the ECML/PKDD Discovery Challenge [45] were achieved by using host level link features only, outperforming those who used all features [2].

In this paper we address not just the quality but also the computational efficiency. Earlier lightweight classifiers include Webb et al. [64] describing a procedure based solely on the HTTP session information. Unfortunately they only measure precision, recall and F-measure that are hard to compare with later results on Web spam that use AUC. In fact the F and similar measures greatly depend on the classification threshold and hence make comparison less stable and for this reason they are not used starting with the Web Spam Challenge 2008. Furthermore in [64] the IP address is a key feature that is trivially incorporated in the DC2010 data set by placing all hosts from the same IP address into the same training or testing set. The intuition is that if an IP address contains spam hosts, all hosts from that IP address are likely to be spam and should be immediately manually checked and excluded from further consideration.

The rest of this paper is organized as follows. In Section 2 we describe the data sets used in this paper. We give an overview of temporal features for spam detection and propose new temporal link similarity based ones in Section 3. In Section 4 we describe our classification framework. The results of the experiments to classify WEBSPAM-UK2007, ClueWeb09 and DC2010 can be found in Section 5. The computational resource needs of various feature sets are summarized in Section 6.

2 Data Sets

In this paper we use three data sets, WEBSPAM-UK2007 of the Web Spam Challenge 2008 [18], the Waterloo Spam Rankings for ClueWeb09, and DC2010 created for the ECML/PKDD Discovery Challenge 2010 on Web Quality. We only give a brief summary of the first data set described well in [18, 22] and the second in [38], however, we describe the third one in more detail in Section 2.3. Also we compare the amount of spam in the data sets.

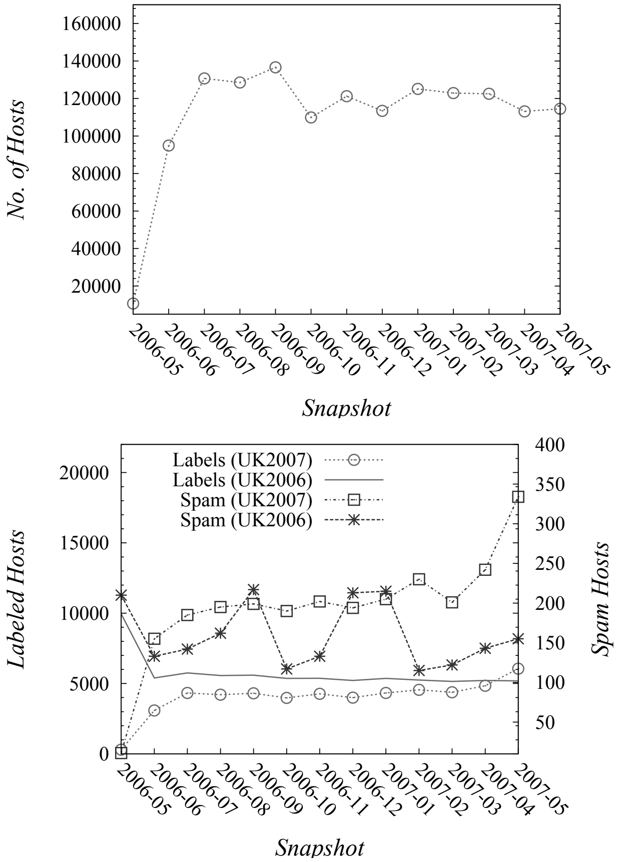


Figure 1: The number of total and labeled hosts in the 13 UK snapshots. We indicate the number of positive and negative labels separate for the WEBSPAM-UK2006 and WEBSPAM-UK2007 label sets.

2.1 Web Spam Challenge 2008: WEBSPAM-UK2007

The Web Spam Challenge was first organized in 2007 over the WEBSPAM-UK2006 data set. The last Challenge over the WEBSPAM-UK2007 set was held in conjunction with AIRWeb 2008 [18]. The Web Spam Challenge 2008 best result [44] achieved an AUC of 0.85 by also using ensemble undersampling [23]. They trained a bagged classifier on the standard content-based and link-based features published by the organizers of the Web Spam Challenge 2008 and on custom host-graph based features, using the ERUS strategy for class-inbalance learning. For earlier challenges, best performances were achieved by a semi-supervised version of SVM [1] and text compression [28]. Best results either used bag of words vectors or the so-called “public” feature sets of [20].

We extended the WEBSPAM-UK2007 data set with 13 `.uk` snapshots pro-

Label Set	**Instances**	**%Positive**
Training	4000	5.95%
Testing	2053	4.68%

Table 1: Summary of label sets for Web Spam Challenge 2008.

vided by the Laboratory for Web Algorithmics of the Università degli studi di Milano. We use the training and testing labels of the Web Spam Challenge 2008, as summarized in Table 1. In order to prepare a temporal collection, we extracted maximum 400 pages per site from the original crawls. The last 12 of the above `.uk` snapshots were analyzed by Bordino et al. [12] who observe a relative low URL but high host overlap[2]. The first snapshot (2006-05) that is identical to WEBSPAM-UK2006 was chosen to be left out from their experiment since it was provided by a different crawl strategy. We observed that in the last eight snapshots the number of hosts have stabilized in the sample and these snapshots have roughly the same amount of labeled hosts as seen in Fig. 1. From now on we restrict attention to the aforementioned subset of the snapshots and the WEBSPAM-UK2007 labels only.

2.2 The Waterloo Spam Rankings for ClueWeb09

The English part of ClueWeb09 consist of approximately 20M domains and 500M pages. For Web spam labels we used the Waterloo Spam Rankings [29]. While the Waterloo Spam Rankings contain negative training instances as well, we extended the negative labels with the set of the Open Directory Project (ODP) hosts. We used 50% split for training and testing.

We labeled hosts in both the `.pt` crawl and ClueWeb09 by top-level ODP categories using links extracted from topic subtrees in the directory. Out of all labeled hosts, 642643 received a unique label. Because certain sites (e.g., `bbc.co.uk`) may belong to even all 14 top-level English categories, we discarded the labels of 18734 hosts with multiple labels to simplify the multi-label task. As Bordino et al. [13] indicate, multitopical hosts are often associated to poor quality sites and spam as another reason why their labels may mislead the classification process. The resulting distribution of labels is shown in Table 2.

2.3 Discovery Challenge 2010: DC2010

The Discovery Challenge was organized over DC2010, a new data set that we describe in more detail next. DC2010 is a large collection of annotated Web hosts labeled by the Hungarian Academy of Sciences (English documents), Internet Memory Foundation (French) and L3S Hannover (German). The base data is a set of 23M pages in 190K hosts in the `.eu` domain crawled by the Internet Memory Foundation in early 2010.

[2] The dataset can be downloaded from: `http://law.di.unimi.it/datasets.php`

Category	**No. of Hosts**	**% of Labeled Hosts**
spam	439	0.07%
Arts	97355	15.1%
Business	193678	30.1%
Computers	66159	10.3%
Recreation	65594	10.2%
Science	43317	6.7%
Society	122084	19%
Sports	54456	8.5%

Table 2: Number of positive ClueWeb09 host labels for spam and the ODP categories.

	UK2006	**UK2007**	**ClueWeb09**	**DC2010**			
				en	de	fr	all
Hosts	10 660	114 529	500,000	61 703	29 758	7 888	190 000
Spam	19.8%	5.3%	unknown	8.5% of valid labels; 5% of all in large domains.			

Table 3: Fraction of Spam in WEBSPAM-UK2006, UK2007, ClueWeb09 and DC2010. Note that three languages English, German and French were selected for labeling DC2010, although Polish and Dutch language hosts constitute a larger fraction than the French. Since to our best knowledge, no systematic random sample was labeled for ClueWeb09, the number 439 of labeled spam hosts is not representative for the collection.

The labels extend the scope of previous data sets on Web Spam in that, in addition to sites labeled spam, we included manual classification for genre into five categories Editorial, Commercial, Educational, Discussion and Personal as well as trust, factuality and bias as three aspects of quality. Spam label is exclusive since no other assessment was made for spam. However other labels are non-exclusive and hence define nine binary classification problems. We consider no multi-class tasks in this paper. Assessor instructions are for example summarized in [62], a paper concentrating on quality labels.

In Table 3, we summarize the amount of spam in the DC2010 data set in comparison with the Web Spam Challenge data sets. This amount is well-defined for the latter data sets by the way they were prepared for the Web Spam Challenge participants. However for DC2010, this figure may be defined in several ways. First of all, when creating the DC2010 labels, eventually we considered domains with or without a `www.` prefix the same such as `www.domain.eu` vs. `domain.eu`. However in our initial sampling procedure we considered them as two different hosts and merged them after verifying that the labels of the two versions were identical. Also, several domains consist of a single redirection page to another domain and we counted these domains, too. Finally, a large fraction of spam is easy to spot and can be manually removed. As an example of many

Count	IP address	Comment
3544	80.67.22.146	spam farm *-palace.eu
3198	78.159.114.140	spam farm *auts.eu
1374	62.58.108.214	blogactiv.eu
1109	91.204.162.15	spam farm x-mp3.eu
1070	91.213.160.26	spam farm a-COUNTRY.eu
936	81.89.48.82	autobazar.eu
430	78.46.101.76	spam farm 77k.eu and 20+ domains
402	89.185.253.73	spam farm mp3-stazeni-zdarma.eu

Table 4: Selection of IP addresses with many subdomains in the DC2010 data set.

Label	Group	Yes	Maybe	No
Spam	*Spam*	423		4 982
News/Editorial	*Genre*	191		4 791
Commercial		2 064		2 918
Educational		1 791		3 191
Discussion		259		4 724
Personal-Leisure		1 118		3 864
Non-Neutrality	*Quality*	19	216	3 778
Bias		62		3 880
Dis-Trustiness		26	201	3 786

Table 5: Distribution of assessor labels in the DC2010 data set.

hosts on same IP, we include a labeled sample from DC2010, that itself contains over 10,000 spam domains in Table 4. These hosts were identified by manually looking at the IP addresses that serve the largest number of domain names. Thus our sample is biased and obtaining an estimate of the spam fraction is nontrivial, as indicated in Table 3.

The distribution of labels for the nine categories with more than 1% positive samples (spam, news, commercial, educational, discussion, personal, neutral, biased, trusted) is given in Table 5. For Neutrality and Trust the strong negative categories have low frequency and hence we fused them with the intermediate negative (maybe) category for the training and testing labels.

The Discovery Challenge 2010 best result [58] achieved an AUC of 0.83 for spam classification while the overall winner [45] was able to classify a number of quality components at an average AUC of 0.80. As for the technologies, bag of words representation variants proved to be very strong for the English collection while only language independent features were used for German and French. The applicability of dictionaries and cross-lingual technologies remains open.

New to the construction of the DC2010 training and test set is the handling of hosts from the same domain and IP address. Since no IP address and domain was allowed to be split between training and testing, we might have to reconsider the applicability of propagation [48, 66] and graph stacking [54]. The Web Spam Challenge data sets were labeled by uniform random sampling and graph stacking appeared to be efficient in several results [22] including our prior work [30]. The applicability of graph stacking remains however unclear for the DC2010 data set. Certain teams used some of these methods but reported no improvement [2].

3 Temporal Features for Spam Detection

Spammers often create bursts in linkage and content: they may add thousands or even millions of machine generated links to pages that they want to promote [61] that they again very quickly regenerate for another target or remove if blacklisted by search engines. Therefore changes in both content and linkage may characterize spam pages.

Recently the evolution of the Web has attracted interest in defining features, signals for ranking [34] and spam filtering [61, 55, 31, 52, 37]. The earliest results investigate the changes of Web content with the primary interest of keeping a search engine index up-to-date [25, 26]. The decay of Web pages and links and its consequences on ranking are discussed in [4, 35]. One main goal of Boldi et al. [11] who collected the `.uk` crawl snapshots also used in our experiments was the efficient handling of time-aware graphs. Closest to our temporal features is the investigation of host overlap, deletion and content dynamics in the same data set by Bordino et al. [12].

Perhaps the first result on the applicability of temporal features for Web spam filtering is due to Shen et al. [61] who compare pairs of crawl snapshots and define features based on the link growth and death rate. However by extending their ideas to consider multi-step neighborhood, we are able to define a very strong feature set that can be computed by the Monte Carlo estimation of Fogaras and Rácz [41]. Another result defines features based on the change of the content [31] who obtain page history from the Wayback Machine.

For calculating the temporal link-based features we use the host level graph. As observed in [12], pages are much more unstable over time compared to hosts. Note that page-level fluctuations may simply result from the sequence the crawler visited the pages and not necessarily reflect real changes. The inherent noise of the crawling procedure and problems with URL canonization [5] rule out the applicability of features based on the change of page-level linkage.

3.1 Linkage Change

In this section we describe link-based temporal features that capture the extent and nature of linkage change. These features can be extracted from either the page or the host level graph where the latter has a directed link from host a to host b if there is a link from a page of a to a page of b.

The starting point of our new features is the observation of [61] that the in-link growth and death rate and change of clustering coefficient characterize the evolution patterns of spam pages. We extend these features for the multi-step neighborhood in the same way as PageRank extends the in-degree. The *ℓ-step neighborhood* of page v is the set of pages reachable from v over a path of length at most ℓ. The ℓ-step neighborhood of a host can be defined similarly over the host graph.

We argue that the changes in the multi-step neighborhood of a page should be more indicative of the spam or honest nature of the page than its single-step neighborhood because spam pages are mostly referred to by spam pages [21], and spam pages can be characterized by larger change of linkage when compared to honest pages [61].

In the following we review the features related to linkage growth and death from [61] in Section 3.1.1, then we introduce new features based on the similarity of the multi-step neighborhood of a page or host. We show how the XJaccard and PSimRank similarity measure can be used for capturing linkage change in Section 3.1.3 and Section 3.1.4, respectively.

3.1.1 Change Rate of In-links and Out-links

We compute the following features introduced by Shen et al. [61] on the host level for a node a for graph instances from time t_0 and t_1. We let $G(t)$ denote the graph instance at time t and $I^{(t)}(a)$, $\Gamma^{(t)}(a)$ denote the set of in and out-links of node a at time t, respectively.

- In-link death (IDR) and growth rate (IGR):

$$\mathrm{IDR}(a) = \frac{|I^{(t_0)}(a)| - |I^{(t_0)}(a) \cap I^{(t_1)}(a)|}{|I^{(t_0)}(a)|}$$

$$\mathrm{IGR}(a) = \frac{|I^{(t_1)}(a)| - |I^{(t_0)}(a) \cap I^{(t_1)}(a)|}{|I^{(t_0)}(a)|}$$

- Out-link death and growth rates (ODR, OGR): the above features calculated for out-links;
- Mean and variance of IDR, IGR, ODR and OGR across in-neighbors of a host (IDRMean, IDRVar, etc.);
- Change rate of the clustering coefficient (CRCC), i.e. the fraction of linked hosts within those pointed by pairs of edges from the same host:

$$CC(a,t) = \frac{|\{(b,c) \in G(t) | b, c \in \Gamma^{(t)}(a)|}{|\Gamma^{(t)}(a)|}$$

$$CRCC(a) = \frac{CC(a,t_1) - CC(a,t_0)}{CC(a,t_0)}$$

- Derivative features such as the ratio and product of the in and out-link rates, means and variances. We list the in-link derivatives; out-link ones are defined similarly:

IGR·IDR, IGR/IDR, IGRMean/IGR, IGRVar/IGR, IDRMean/IDR, IDRVar/IDR, IGRMean · IDRMean, IGRMean/IDRMean, IGRVar · IDRVar, IGRVar/IDRVar.

3.1.2 Self-Similarity Along Time

In the next sections we introduce new linkage change features based on multi-step graph similarity measures that in some sense generalize the single-step neighborhood change features of the previous section. We characterize the change of the multi-step neighborhood of a node by defining the similarity of a single node *across* snapshots instead of two nodes within a single graph instance. The basic idea is that, for each node, we measure its similarity to itself in two identically labeled graphs representing two consecutive points of time. This enables us to measure the linkage change occurring in the observed time interval using ordinary graph similarity metrics.

First we describe our new contribution, the extension of two graph similarity measures, XJaccard and PSimRank [41] to capture temporal change; moreover, we argue why SimRank [51] is inappropriate for constructing temporal features.

SimRank of a pair of nodes u and v is defined recursively as the average similarity of the neighbors of u and v:

$$\begin{aligned} \mathrm{Sim}_{\ell+1}(u,v) &= 0, \text{ if } I(u) \text{ or } I(v) \text{ is empty;} \\ \mathrm{Sim}_{\ell+1}(u,v) &= 1, \text{ if } u = v; \\ \mathrm{Sim}_{\ell+1}(u,v) &= \frac{c}{|I(u)||I(v)|} \sum_{\substack{v'\in I(v)\\ u'\in I(u)}} \mathrm{Sim}_{\ell}(u',v'), \end{aligned} \tag{1}$$

where $I(x)$ denotes the set of vertices linking to x and $c \in (0,1)$ is a decay factor. In order to apply SimRank for similarity of a node v between two snapshots t_0 and t_1, we apply (2) so that v' and u' are taken from different snapshots.

Next we describe a known deficiency of SimRank in its original definition that rules out its applicability for temporal analysis. First we give the example for the single graph SimRank. Consider a bipartite graph with k nodes pointing all to another two u and v. In this graph there are no directed paths of length more than one and hence the Sim values can be computed in a single iteration. Counter-intuitively, we get $\mathrm{Sim}(u,v) = c/k$, i.e. the larger the cocitation of u and v, the smaller their SimRank value. The reason is that the more the number of in-neighbors, the more likely is that a pair of random neighbors will be different.

While the example of the misbehavior for SimRank is somewhat artificial in the single-snapshot case, next we show that this phenomenon almost always happens if we consider the similarity of a single node v across two snapshots. If there is no change at all in the neighborhood of node v between the two snapshots, we expect the Sim value to be maximal. However the situation is identical to the bipartite graph case and Sim will be inversely proportional to the number of out-links.

3.1.3 Extended Jaccard Similarity Along Time

Our first definition of similarity is based on the extension of the Jaccard coefficient in a similar way XJaccard is defined in [41]. The Jaccard similarity of a page or host v across two snapshots t_0 and t_1 is defined by the overlap of its neighborhood in the two snapshots, $\Gamma^{(t_0)}(v)$ and $\Gamma^{(t_1)}(v)$ as

$$\mathsf{Jac}^{(t_0,t_1)}(v) = \frac{|\Gamma^{(t_0)}(v) \cap \Gamma^{(t_1)}(v)|}{|\Gamma^{(t_0)}(v) \cup \Gamma^{(t_1)}(v)|}$$

The *extended Jaccard coefficient, XJaccard* for length ℓ of a page or host is defined via the notion of the neighborhood $\Gamma_k^{(t)}(v)$ at distance exactly k as

$$\mathsf{XJac}_\ell^{(t_0,t_1)}(v) = \sum_{k=1}^{\ell} \frac{|\Gamma_k^{(t_0)}(v) \cap \Gamma_k^{(t_1)}(v)|}{|\Gamma_k^{(t_0)}(v) \cup \Gamma_k^{(t_1)}(v)|} \cdot c^k(1-c),$$

where c is a decay factor.

The XJac values can be approximated by the min-hash fingerprinting technique for Jaccard coefficients [15], as described in Algorithm 3 of [41]. The fingerprint generation algorithm has to be repeated for each graph snapshot, with the same set of independent random permutations.

We generate temporal features based on the XJac values for four length values $\ell = 1 \ldots 4$. We also repeat the computation on the transposed graph, i.e. replacing out-links $\Gamma^{(t)}(v)$ by in-links $I^{(t)}(v)$. As suggested in [41], we set the decay factor $c = 0.1$ as this is the value where, in their experiments, XJaccard yields best average quality for similarity prediction.

Similar to [61], we also calculate the mean and variance $\mathsf{XJac}^{(t_0,t_1)}{}_\ell(w)$ of the neighbors w for each node v. The following derived features are also calculated:

- similarity at path length $\ell = 2, 3, 4$ divided by similarity at path length $\ell - 1$, and the logarithm of these;
- logarithm of the minimum, maximum, and average of the similarity at path length $\ell = 2, 3, 4$ divided by the similarity at path length $\ell - 1$.

3.1.4 PSimRank Along Time

Next we define similarity over time based on PSimRank, a SimRank variant defined in [41] that can be applied similar to XJaccard in the previous section. As we saw in Section 3.1.2, SimRank is inappropriate for measuring linkage change in time. In the terminology of the previous subsection, the reason is that path fingerprints will be unlikely to meet in a large neighborhood and SimRank values will be low even if there is completely no change in time.

We solve the deficiency of SimRank by allowing the random walks to meet with higher probability when they are close to each other: a pair of random walks at vertices u', v' will advance to the same vertex (i.e., meet in one step) with probability of the Jaccard coefficient $\frac{|I(u') \cap I(v')|}{|I(u') \cup I(v')|}$ of their in-neighborhood $I(u')$ and $I(v')$.

The random walk procedure corresponding to PSimRank along with a fingerprint generation algorithm is defined in [41].

For the temporal version, we choose independent random permutations σ_ℓ on the hosts for each step ℓ. In step ℓ if the random walk from vertex u is at u', it will step to the in-neighbor with smallest index given by the permutation σ_ℓ in each graph snapshot.

Temporal features are derived from the PSimRank similarity measure very much the same way as for XJaccard, for four length values $\ell = 1 \dots 4$. We also repeat the computation on the transposed graph, i.e. replacing out-links $\Gamma^{(t)}(v)$ by in-links $I^{(t)}(v)$. As suggested in [41], we set the decay factor $c = 0.15$ as this is the value where, in their experiments, PSimRank yields best average quality for similarity prediction. Additionally, we calculate the mean and variance PSimRank(w) of the neighbors w for each node v and derived features as for XJaccard.

3.2 Content and its Change

The content of Web pages can be deployed in content classification either via statistical features such as entropy [59] or via term weight vectors [67, 31]. Some of the more complex features that we do not consider in this work include language modeling [3].

In this section we focus on capturing term-level changes over time. For each target site and crawl snapshot, we collect all the available HTML pages and represent the site as the bag-of-words union of all of their content. We tokenize content using the ICU library[3], remove stop words[4] and stem using Porter's method.

We treat the resulting term list as the virtual document for a given site at a point of time. As our vocabulary we use the most frequent 10,000 terms found in at least 10% and at most 50% of the virtual documents.

To measure the importance of each term i in a virtual document d at time snapshot T, we use the BM25 weighting [60]:

$$t_{i,d}^{(T)} = \mathrm{IDF}_i^{(T)} \cdot \frac{(k_1 + 1)\mathrm{tf}_{i,d}^{(T)}}{K + \mathrm{tf}_{i,d}^{(T)}}$$

where $\mathrm{tf}_{i,d}^{(T)}$ is the number of occurrences of term i in document d and $\mathrm{IDF}_i^{(T)}$ is the inverse document frequency (Robertson-Spärck Jones weight) for the term at time T. The length normalized constant K is specified as

$$k_1((1 - b) + b \times \mathrm{dl}^{(T)}/\mathrm{avdl}^{(T)})$$

such that $dl^{(T)}$ and $avdl^{(T)}$ denote the virtual document length and the average length over all virtual documents at time T, respectively. Finally

$$\mathrm{IDF}^{(T)} = log \frac{N - n^{(T)} + 0.5}{n^{(T)} + 0.5}$$

[3] http://icu-project.org/

[4] http://www.lextek.com/manuals/onix/stopwords1.html

where N denotes the total number of virtual documents and $n^{(T)}$ is the number of virtual documents containing term i. Note that we keep N independent of T and hence if document d does not exist at T, we consider all $\mathrm{tf}_{i,d}^{(T)} = 0$.

By using the term vectors as above, we calculate the temporal content features described in [31] in the following five groups.

- **Ave:** Average BM25 score of term i over the $T_{\max}$ snapshots:

$$\mathrm{Ave}_{i,d} = \frac{1}{T_{\max}} \cdot \sum_{T=1}^{T_{\max}} t_{i,d}^{(T)}$$

- **AveDiff:** Mean difference between temporally successive term weight scores:

$$\mathrm{AveDiff}_{i,d} = \frac{1}{T_{\max} - 1} \cdot \sum_{T=1}^{T_{\max}-1} |t_{i,d}^{(T+1)} - t_{i,d}^{(T)}|$$

- **Dev:** Variance of term weight vectors at all time points:

$$\mathrm{Dev}_{i,d} = \frac{1}{T_{\max} - 1} \cdot \sum_{T=1}^{T_{\max}} (t_{i,d}^{(T)} - \mathrm{Ave}_{i,d})^2$$

- **DevDiff:** Variance of term weight vector differences of temporally successive virtual documents:

$$\mathrm{DevDiff}_{i,d} = \frac{1}{T_{\max} - 2} \cdot \sum_{T=1}^{T_{\max}-1} (|t_{i,d}^{(T+1)} - t_{i,d}^{(T)}| - \mathrm{AveDiff}_{i,d})^2$$

- **Decay:** Weighted sum of temporally successive term weight vectors with exponentially decaying weight. The base of the exponential function, the *decay rate* is denoted by λ. **Decay** is defined as follows:

$$\mathrm{Decay}_{i,d} = \sum_{T=1}^{T_{\max}} \lambda e^{\lambda(T_{\max}-T)} t_{i,d}^{(T)}$$

4 Classification Framework

For the purposes of our experiments we computed all the public Web Spam Challenge content and link features of [20]. We built a classifier ensemble by splitting features into related sets and for each we use a collection of classifiers that fit the data type and scale. These classifiers were then combined by ensemble selection. We used the classifier implementations of the machine learning toolkit Weka [65].

Ensemble selection is an overproduce and choose method allowing to use large collections of diverse classifiers [17]. Its advantages over previously published methods [16] include optimization to any performance metric and refinements to prevent overfitting, the latter being unarguably important when more classifiers are available for selection. The motivation for using ensemble selection is that recently this particular ensemble method gained more attention thanks to the winners of KDD Cup 2009 [57]. In our experiments [38] ensemble selection performed significantly better than other classifier combination methods used for Web spam detection in the literature, such as log-odds based averaging [56] and bagging.

In the context of combining classifiers for Web classification, to our best knowledge, ensemble selection has not been applied yet. Previously, only simple methods that combine the predictions of SVM or decision tree classifiers through logistic regression or random forest have been used [28]. We believe that the ability to combine a large number of classifiers while preventing overfitting makes ensemble selection an ideal candidate for Web classification, since it allows us to use a large number of features and learn different aspects of the training data at the same time. Instead of tuning various parameters of different classifiers, we can concentrate on finding powerful features and selecting the main classifier models which we believe to be able to capture the differences between the classes to be distinguished.

We used the ensemble selection implementation of Weka [65] for performing the experiments. Weka's implementation supports the proven strategies to avoid overfitting such as model bagging, sort initialization and selection with replacement. We allow Weka to use all available models in the library for greedy sort initialization and use 5-fold embedded cross-validation during ensemble training and building. We set AUC as the target metric to optimize for and run 100 iterations of the hillclimbing algorithm.

We mention that we have to be careful with treating missing feature values. Since the temporal features are based on at least two snapshots, for a site that appears only in the last one, all temporal features have missing value. For classifiers that are unable to treat missing values we define default values depending on the type of the feature.

4.1 Learning Methods

We use the following models in our ensemble: bagged and boosted decision trees, logistic regression, naive Bayes and variants of random forests. For most classes of features we use all classifiers and let selection choose the best ones. The exception is static and temporal term vector based features where, due to the very large number of features, we may only use Random Forest and SVM. We train our models as follows.

Bagged LogitBoost: we do 10 iterations of bagging and vary the number of iterations from 2 to 64 in multiples of two for LogitBoost.

Decision Trees: we generate J48 decision trees by varying the splitting criterion, pruning options and use either Laplacian smoothing or no smoothing at all.

Bagged Cost-sensitive Decision Trees: we generate J48 decision trees with default parameters but vary the cost sensitivity for false positives in steps of 10 from 10 to 300. We do the same number of iterations of bagging as for LogitBoost models.

Logistic Regression: we use a regularized model varying the ridge parameter between 10^{-8} to 10^4 by factors of 10. We normalize features to have mean 0 and standard deviation 1.

Random Forests: we use FastRandomForest [39] instead of the native Weka implementation for faster computation. The forests have 250 trees and, as suggested in [14], the number of features considered at each split is $s/2$, s, $2s$, $4s$ and $8s$, where s is the square root of the total number of features available.

Naive Bayes: we allow Weka to model continuous features either as a single normal or with kernel estimation, or we let it discretize them with supervised discretization.

5 Results and Discussion

In this section we describe the various ensembles we built and measure their performance[5]. We compare feature sets by using the same learning methods described in Section 4 while varying the subset of features available for each of the classifier instances when training and combining these classifiers using ensemble selection. We also concentrate on the value of temporal information for Web spam detection. As our goal is to explore the computational cost vs. classification performance tradeoff, we will describe the resource needs for various features in detail in Section 6.

For training and testing we use the official Web Spam Challenge 2008 training and test sets [20]. As it can be seen in Table 1 these show considerable class imbalance which makes the classification problem harder. For DC2010 we also use the official training set as described in Table 5. For ClueWeb09 we used a 50% random split.

To make it easy to compare our results to previous results, we cite the Web Spam Challenge 2008 and Discovery Challenge 2010 winner's performance in the summary tables next. For ClueWeb09 the only previous evaluation is in terms of TREC retrieval performance [29] that we cannot directly compare here.

5.1 Content-only Ensemble

We build three different ensembles over the content-only features in order to assess performance by completely eliminating linkage information. The feature sets available for these ensembles are the following:

- (A) Public content [59, 22] features without any link based information. Features for the page with maximum PageRank in the host are not used to save the PageRank computation. Corpus precision, the fraction of words in a page that is corpuswise frequent and corpus recall, the fraction

[5] The exact classifier model specification files used for Weka and the data files used for the experiments are available upon request from the authors.

of corpuswise frequent terms in the page are not used either since they require global information from the corpus.

- (Aa) The tiniest feature set of 24 features from (A): query precision and query recall defined similar to corpus precision and recall but based on popular terms from a proprietary query log[6] instead of the entire corpus. A very strong feature set based on the intuition that spammers use terms that make up popular queries.
- (B) The full public content feature set [22], including features for the maximum PageRank page of the host.
- Feature set (B) plus a bag of words representation derived from the BM25 [60] term weighting scheme.

Table 6 presents the performance comparison of ensembles built using either of the above feature sets. The DC2010 and ClueWeb09 detailed results are in Table 8 and Table 9, respectively. Performance is given in AUC for all data sets.

Feature Set	Number of Features	UK2007	DC2010	ClueWeb09
Content (A)	74	0.859	0.757	0.829
Content (Aa)	24	0.841	0.726	0.635
Content (B)	96	0.879	0.799	0.827
BM25 + (B)	10096	**0.893**	**0.891**	**0.870**
Challenge best	-	0.852	0.830	-

Table 6: AUC value of spam ensembles built from content based features.

Surprisingly, with the small (Aa) feature set of only 24 features a performance only 1% worse than that of the Web Spam Challenge 2008 winner can be achieved who employed more sophisticated methods to get their result. By using all the available content based features without linkage information, we get roughly the same performance as the best which have been reported on our data set so far. However this achievement can be rather attributed to the better machine learning techniques used than the feature set itself since the features used for this particular measurement were already publicly accessible at the time of the Web Spam Challenge 2008.

As it can be seen in Table 6 relative performance of content based features over different corpora varies a lot. In case of DC2010 and ClueWeb09 the small (Aa) feature set achieves much worse result than the largest feature set having best performance for all data sets. The fact that the content (A, Aa, B) and link (Table 7) performances are always better for UK2007 might be explained

[6] A summary is available as part of our data release at `https://dms.sztaki.hu/sites/dms.sztaki.hu/files/download/2013/enpt-queries.txt.gz`.

by the fact that the UK2007 training and testing sets were produced by random sampling without considering domain boundaries. Hence in a large domain with many subdomains, part of the hosts belong to the training and part to the testing set with very similar distribution. This advantage disappears for the BM25 features.

5.2 Full Ensemble

Feature Set	Number of Features	UK2007	DC2010	ClueWeb09
Public link-based [7]	177	0.759	0.587	0.806
All combined	10 273	**0.902**	0.885	0.876

Table 7: Performance of ensembles built on link based and all features.

Results of the ensemble incorporating all the previous classifiers is seen in Table 7. The DC2010 detailed results are in Table 8. Overall, we observe that BM25 is a very strong feature set that may even be used itself for a lightweight classifier. On the other hand, link features add little to quality and the gains apparently diminish for DC2010, likely due to the fact that the same domain and IP address is not split between training and testing.

The best Web Spam Challenge 2008 participant [44] reaches an AUC of 0.85 while for DC2010, the best spam classification AUC of [58] is 0.83. We outperform these results by a large margin.

For DC2010 we also show detailed performance for nine attributes in Table 8, averaged in three groups: spam, genre and quality (as in Table 5). Findings are similar: with BM25 domination, part or all of the content features slightly increase the performance. Results for the quality attributes and in particular for trust are very low. Classification for these aspects remains a challenging task for the future.

For ClueWeb09 detailed performance for selected ODP categories can be seen in Table 9. Identically to DC2010 results BM25 features provide the best classification performance. However, combinations with other feature sets yield gains only for spam classification. For the ODP classification tasks linkage information does not help in general: the content based feature set has roughly the same performance with or without page-level linkage information, and combining with the link based feature set does not improve performance notably in most labeling tasks.

5.3 Temporal Link Ensembles

First, we compare the temporal link features proposed in Section 3.1 with those published earlier [61]. Then, we build ensembles that combine the temporal with

Feature Set	spam	genre **average**	quality **average**	**average**
Public link-based [7]	0.655	0.614	0.519	0.587
Content (A)	0.757	0.713	0.540	0.660
Content (Aa)	0.726	0.662	0.558	0.634
Content (B)	0.799	0.735	0.512	0.668
BM25	**0.876**	**0.805**	**0.584**	**0.739**
Public link-based + (B)	0.812	0.731	0.518	0.669
BM25 + (A)	0.872	**0.816**	0.580	**0.754**
BM25 + (B)	**0.891**	0.810	**0.612**	0.744
All combined	0.885	0.813	0.553	0.734

Table 8: Performance over the DC2010 labels in terms of AUC.

Feature Set	spam	Arts	Business	Computers	Recreation	Science	Society	Sports	**ODP average**
Link [7]	.806	.569	.593	.591	.532	.624	.540	.504	.595
Content (A)	.829	.676	.726	.632	.669	.720	.639	.673	.695
Content (Aa)	.635	.508	.524	.554	.487	.558	.502	.522	.536
Content (B)	.827	.673	.727	.634	.670	.720	.629	.674	.694
BM25	.845	**.913**	**.890**	**.931**	**.907**	**.883**	**.915**	**.959**	**.914**
Link + (B)	.848	.675	.731	.646	.669	.727	.631	.669	.699
BM25 + (A)	.871	.895	.881	.896	.879	.851	.904	.935	.892
BM25 + (B)	.869	.895	.881	.898	.892	.850	.906	.934	.894
All combined	**.876**	.896	.883	.898	.892	.852	.905	.936	.895

Table 9: Performance over the ClueWeb09 labels in terms of AUC.

the public link-based features described by [7]. The results are summarized in Table 10. Note that all experiments in this section and Section 5.4 were carried out on the WEBSPAM-UK2007 data set.

As these measurements show, our proposed graph similarity based features successfully extend the growth and death rate based ones by achieving higher accuracy, improving AUC by 1.3%. However, by adding temporal to static link-based features we get only marginally better ensemble performance.

To rank the link-based feature sets by their contribution in the ensemble, we build classifier models on the three separate feature subsets (public link-based, growth/death rate based and graph similarity based features, respectively) and let ensemble selection combine them. This restricted combination results in a slightly worse AUC of 0.762. By calculating the total weight contribution,

Section	**Feature Set**	**No. of Features**	**AUC**
3.1.1	Growth/death rates	29	0.617
3.1.3-4	XJaccard + PSimRank	63	0.625
	Public link-based [7]	176	0.765
3.1.1	Public + growth/death rates	205	0.758
3.1.3-4	Public + XJaccard + PSimRank	239	**0.769**
	All link-based	268	0.765
	WSC 2008 Winner	-	0.852

Table 10: Performance of ensembles built on link-based features.

we get the following ranked list (weight contribution showed in parenthesis): public link-based (60.8%), graph similarity based (21.5%), growth/death rate based (17.7%). This ranking also supports the findings presented in Table 10 that graph similarity based temporal link-based features should be combined with public link-based features if temporal link-based features are used.

To separate the effect of ensemble selection on the performance of temporal link-based feature sets we repeat the experiments with bagged cost-sensitive decision trees only, a model reported to be effective for web spam classification [59]. The results for these experiments are shown in Table 11.

Section	**Feature Set**	**No. of Features**	**AUC**
3.1.1	Growth/death rates	29	0.605
3.1.3	XJaccard	42	0.626
3.1.4	PSimRank	21	0.593
3.1.3-4	XJaccard + PSimRank	63	0.610
	Public link-based [7]	176	**0.731**
3.1.1	Public + growth/death rates	205	0.696
3.1.3-4	Public + XJaccard + PSimRank	239	*0.710*
	All link-based	268	0.707
	WSC 2008 Winner	-	0.852

Table 11: Performance of bagged cost-sensitive decision trees trained on link-based features.

As it can be seen in Table 11, when using bagged cost-sensitive decision trees, our proposed temporal link-based similarity features achieve 3.5% better performance than the growth/death rate based features published earlier.

When comparing results in Table 11 and in Table 10 we can see that ensemble selection i) significantly improves accuracy (as expected) and ii) diminishes the performance advantage achievable by the proposed temporal link-based features over the previously published ones.

As evident from Table 11, the proposed PSimRank based temporal features perform roughly the same as the growth and death rate based ones while the XJaccard based temporal features perform slightly better.

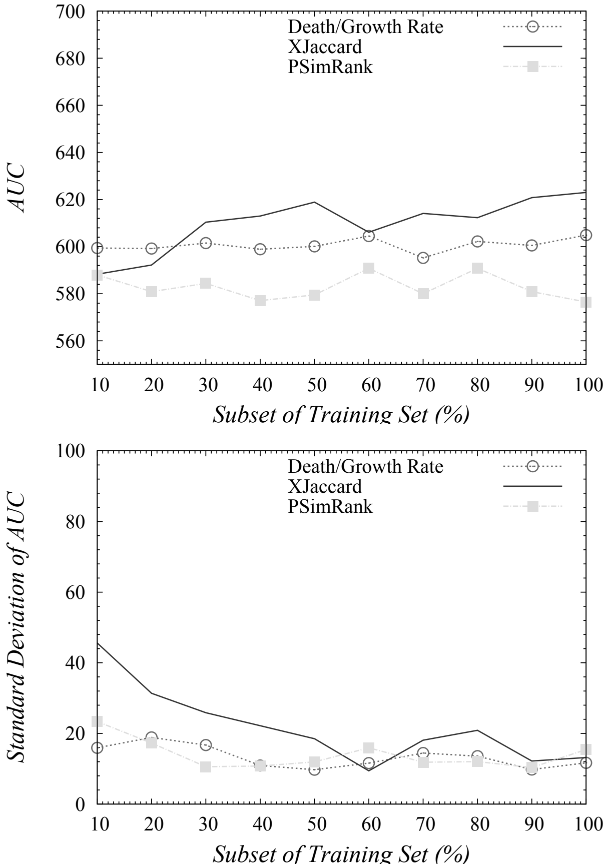


Figure 2: Sensitivity of temporal link-based features. **Top:** AUC values averaged across 10 measurements. **Bottom:** standard deviations of AUC for different training set sizes.

Next we perform sensitivity analysis of the temporal link-based features by using bagged cost-sensitive decision trees. We build 10 different random training samples for each of the possible fractions 10%, 20%, . . . , 100% of all available labels. In Fig. 2 we can see that the growth/death rate based features as well as the PSimRank based features are not sensitive to training set size while the

XJaccard based ones are. That is, even though XJaccard is better in terms of performance than the other two feature sets considered it is more sensitive to the amount of training data used as well.

5.4 Temporal Content Ensembles

We build ensembles based on the temporal content features described in Section 3.2 and their combination themselves, with the static BM25 features, and with the content-based features of [59]. The performance comparison of temporal content-based ensembles is presented in Table 12.

Feature Set	**No. of Features**	**AUC**
Static BM25	10,000	0.736
Ave	10,000	0.749
AveDiff	10,000	0.737
Dev	10,000	0.767
DevDiff	10,000	0.752
Decay	10,000	0.709
Temporal combined	50,000	0.782
Temporal combined + BM25	60,000	**0.789**
Public content-based [59] + temporal	50,096	0.901
All combined	60,096	**0.902**

Table 12: Performance of ensembles built on temporal content-based features.

By combining all the content and link-based features, both temporal and static ones, we train an ensemble which incorporates all the previous classifiers. This combination resulted in an AUC of 0.908 meaning no significant improvement can be achieved with link-based features over the content-based ensemble.

6 Computational Resources

For the experiments we used a 45-node Hadoop cluster of dual core machines with 4GB RAM each as well as multi-core machines with over 40GB RAM. Over this architecture we were able to compute all features, some of which would require excessive resources either when used by a smaller archive or if the collection is larger or if fast classification is required for newly discovered sites during crawl time. Some of the most resource bound features involve the multistep neighborhood in the page level graph that already requires approximation techniques for WEBSPAM-UK2007 [22].

We describe the computational requirements of the features by distinguishing update and batch processing. For batch processing an entire collection is analyzed at once, a procedure that is probably performed only for reasons of research. Update is probably the typical operation for a search engine. For an

Feature Set	Step	Hours	Configuration
Content (A) + BM25	Parsing	36	45 dual core Pentium-D 3.0GHz machines, 4GB RAM, Hadoop 0.21
	Feature generation	36	
	Selection of labeled pages	3	
Link	PageRank	10	5 eight-core Xeon 1.6GHz machines, 40+GB RAM
	Neighborhood	4	
	Local features	1	

Table 13: Processing times and cluster configurations for feature sets over ClueWeb09.

Internet Archive, update is also advantageous as long as it allows fast reaction to sample, classify and block spam from a yet unknown site.

6.1 Batch Processing

The first expensive step involves parsing to create terms and links. The time requirement scales linearly with the number of pages. Since apparently a few hundred page sample of each host suffices for feature generation, the running time is also linear in the number of hosts. For a very large collection such as ClueWeb09, distributed processing may be necessary. Over 45 dual core Pentium-D 3.0GHz machines running Hadoop 0.21, we parsed the uncompressed 9.5TB English part of ClueWeb09 in 36 hours. Additional tasks such as term counting, BM25 or content feature generation fits within the same time frame. If features are generated only a small labeled part of the data, it took us 3 hours to select the appropriate documents and additional processing time was negligible. Processing times are summarized in Table 13.

Host level aggregation allows us to proceed with a much smaller size data. However for aggregation we need to store a large number of partial feature values for all hosts unless we sort the entire collection by host, again by external memory or Map-Reduce sort.

After aggregation, host level features are inexpensive to compute. The following features however remain expensive:

- Page level PageRank. Note that this is required for all content features involving the maximum PageRank page of the host.
- Page level features involving multi-step neighborhood such as neighborhood size at distance k as well as graph similarity.

In order to be able to process graphs of ClueWeb09 scale (4.7 billion nodes and 17 billion edges), we implemented message passing C++ codes. Over a total 30 cores of six Xeon 1.6GHz machines, each with at least 40GB RAM, one PageRank and one Bit Propagation iteration both took approximately one hour while all other, local features completed within one hour.

Training the classifier for a few 100,000 sites can be completed within a day on a single CPU on a commodity machine with 4-16GB RAM; here costs

Configuration	**Number of Hosts**	**Feature Sets**	**Example**	**Expected Accuracy**	**Computation**
Small 1-2 machines	10,000	Content (A) BM25	subset of UK2007	0.80-0.87	Non-distributed
Medium 3-10 machines	100,000	Content (A) BM25, link	DC2010	0.87-0.90	MapReduce and Disk-based e.g. GraphChi
Large 10+ machines	1,000,000	Content (B) BM25, link	ClueWeb09	0.9+	MapReduce and Pregel

Table 14: Sample configurations for Web spam filtering in practice.

strongly depend on the classifier implementation. Our entire classifier ensemble for the labeled WEBSPAM-UK2007 hosts took a few hours to train.

6.2 Incremental Processing

As preprocessing and host level aggregation is linear in the number of hosts, this reduces to a small job for an update. This is especially true if we are able to split the update by sets of hosts; in this case we may even trivially parallelize the procedure.

The only nontrivial content based information is related to document frequencies: both the inverse document frequency term of BM25 [60] and the corpus precision and recall dictionaries may in theory be fully updated when new data is added. We may however approximate by the existing values under the assumption that a small update batch will not affect these values greatly. From time to time however all features beyond (Aa) need a global recomputation step.

The link structure is however nontrivial to update. While incremental algorithms exist to create the graph and to update PageRank type features [32, 33, 53], these algorithms are rather complex and their resource requirements are definitely beyond the scale of a small incremental data.

Incremental processing may have the assumption that no new labels are given, since labeling a few thousand hosts takes time comparable to batch process hundreds of thousands of them. Given the trained classifier, a new site can be classified in seconds right after its feature set is computed.

7 Conclusions

With the illustration over the 100,000 host WEBSPAM-UK2007, the half billion page ClueWeb09, and the 190,000 host DC2010 data sets, we have investigated the tradeoff between feature generation and spam classification accuracy. We observe that more features achieve better performance, however, when combining them with the public link based feature set we get only marginal performance gain. By using the WEBSPAM-UK2007 data along with seven previous monthly snapshots of the `.uk` domain, we have presented a survey of temporal features for Web spam classification. We investigated the performance of link, content and temporal[7] Web spam features with ensemble selection. As practical message, we may conclude that, as seen in Table 14, single machines may compute content and BM25 features for a few 10,000 hosts only. Link features need additional resources and either compressed, disk based or, in the largest configuration, Pregel-like distributed infrastructures.

We proposed graph similarity based temporal features which aim to capture the nature of linkage change of the neighborhoods of hosts. We have shown how to compute these features efficiently on large graphs using a Monte Carlo method. Our features achieve better performance than previously published methods, however, when combining them with the public link-based feature set we get only marginal performance gain.

By our experiments it has turned out that the appropriate choice of the machine learning techniques is probably more important than devising new complex features. We have managed to compile a minimal feature set that can be computed incrementally very quickly to allow to intercept spam at crawl time based on a sample of a new Web site. Sample configurations for Web spam filtering are summarized in Table 14.

Our results open the possibility for spam filtering practice in Internet archives who are mainly concerned about their resource waste and would require fast reacting filters. BM25 based models are suitable even for filtering at crawl time.

Some technologies remain open to be explored. For example, unlike expected, the ECML/PKDD Discovery Challenge 2010 participants did not deploy cross-lingual technologies for handling languages other than English. Some ideas worth exploring include the use of dictionaries to transfer a bag of words based model and the normalization of content features across languages to strengthen the language independence of the content features. The natural language processing based features were not used either, that may help in particular with the challenging quality attributes.

Acknowledgment

To Sebastiano Vigna, Paolo Boldi and Massimo Santini for providing us with the UbiCrawler crawls [10, 11]. In addition to them, also to Ilaria Bordino, Carlos Castillo and Debora Donato for discussions on the WEBSPAM-UK data sets [12].

To the large team of organizers and assessors for the complex labeling process of the DC2010 data set.

[7] The temporal feature data used in our research is available at: `https://datamining.sztaki.hu/en/download/web-spam-resources`

References

[1] J. Abernethy, O. Chapelle, and C. Castillo. WITCH: A New Approach to Web Spam Detection. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[2] L. D. Artem Sokolov, Tanguy Urvoy and O. Ricard. Madspam consortium at the ecml/pkdd discovery challenge 2010. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[3] J. Attenberg and T. Suel. Cleaning search results using term distance features. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 21–24. ACM New York, NY, USA, 2008.

[4] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: Towards an understanding of the web's decay. In *Proceedings of the 13th World Wide Web Conference (WWW)*, pages 328–337. ACM Press, 2004.

[5] Z. Bar-Yossef, I. Keidar, and U. Schonfeld. Do not crawl in the dust: different urls with similar text. *ACM Transactions on the Web (TWEB)*, 3(1):1–31, 2009.

[6] S. Barton. Mignify, a big data refinery built on hbase. In *HBASE CON*, 2012.

[7] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of web spam. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2006.

[8] A. A. Benczúr, M. Erdélyi, J. Masanés, and D. Siklósi. Web spam challenge proposal for filtering in archives. In *AIRWeb '09: Proceedings of the 5th international workshop on Adversarial information retrieval on the web*. ACM Press, 2009.

[9] A. A. Benczúr, D. Siklósi, J. Szabó, I. Bíró, Z. Fekete, M. Kurucz, A. Pereszlényi, S. Rácz, and A. Szabó. Web spam: a survey with vision for the archivist. In *Proc. International Web Archiving Workshop*, 2008.

[10] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: A scalable fully distributed web crawler. *Software: Practice & Experience*, 34(8):721–726, 2004.

[11] P. Boldi, M. Santini, and S. Vigna. A Large Time Aware Web Graph. *SIGIR Forum*, 42, 2008.

[12] I. Bordino, P. Boldi, D. Donato, M. Santini, and S. Vigna. Temporal evolution of the uk web. In *Workshop on Analysis of Dynamic Networks (ICDM-ADN'08)*, 2008.

[13] I. Bordino, D. Donato, and R. Baeza-Yates. Coniunge et impera: Multiple-graph mining for query-log analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 168–183. Springer, 2010.

[14] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[15] A. Z. Broder. On the Resemblance and Containment of Documents. In *Proceedings of the Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29, 1997.

[16] R. Caruana, A. Munson, and A. Niculescu-Mizil. Getting the most out of ensemble selection. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 828–833, Washington, DC, USA, 2006. IEEE Computer Society.

[17] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes. Ensemble selection from libraries of models. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 18, New York, NY, USA, 2004. ACM.

[18] C. Castillo, K. Chellapilla, and L. Denoyer. Web spam challenge 2008. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[19] C. Castillo and B. Davison. *Adversarial web search*, volume 4. Now Publishers Inc, 2011.

[20] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

[21] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. Technical report, DELIS – Dynamically Evolving, Large-Scale Information Systems, 2006.

[22] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, 2007.

[23] N. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

[24] C. Chekuri, M. H. Goldwasser, P. Raghavan, and E. Upfal. Web search using automatic classification. In *Proceedings of the 6th International World Wide Web Conference (WWW)*, San Jose, USA, 1997.

[25] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *The VLDB Journal*, pages 200–209, 2000.

[26] J. Cho and H. Garcia-Molina. Synchronizing a database to improve freshness. In *Proceedings of the International Conference on Management of Data*, pages 117–128, 2000.

[27] E. Convey. Porn sneaks way back on web. *The Boston Herald*, May 1996.

[28] G. Cormack. Content-based Web Spam Detection. In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2007.

[29] G. Cormack, M. Smucker, and C. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465, 2011.

[30] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-Supervised Learning: A Comparative Study for Web Spam and Telephone User Churn. In *Graph Labeling Workshop in conjunction with ECML/PKDD 2007*, 2007.

[31] N. Dai, B. D. Davison, and X. Qi. Looking into the past to better classify web spam. In *AIRWeb '09: Proceedings of the 5th international workshop on Adversarial information retrieval on the web*. ACM Press, 2009.

[32] P. Desikan, N. Pathak, J. Srivastava, and V. Kumar. Incremental page rank computation on evolving graphs. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 1094–1095, New York, NY, USA, 2005. ACM.

[33] P. K. Desikan, N. Pathak, J. Srivastava, and V. Kumar. Divide and conquer approach for efficient pagerank computation. In *ICWE '06: Proceedings of the 6th international conference on Web engineering*, pages 233–240, New York, NY, USA, 2006. ACM.

[34] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, K. Buchner, R. Zhang, C. Liao, and F. Diaz. Towards recency ranking in web search. In *Proc. WSDM*, 2010.

[35] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the web frontier. In *Proceedings of the 13th International World Wide Web Conference (WWW)*, pages 309–318, New York, NY, USA, 2004. ACM Press.

[36] M. Erdélyi and A. A. Benczúr. Temporal analysis for web spam detection: An overview. In *1st International Temporal Web Analytics Workshop (TWAW) in conjunction with the 20th International World Wide Web Conference in Hyderabad, India.* CEUR Workshop Proceedings, 2011.

[37] M. Erdélyi, A. A. Benczúr, J. Masanés, and D. Siklósi. Web spam filtering in internet archives. In *AIRWeb '09: Proceedings of the 5th international workshop on Adversarial information retrieval on the web*. ACM Press, 2009.

[38] M. Erdélyi, A. Garzó, and A. A. Benczúr. Web spam classification: a few features worth more. In *Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality 2011) In conjunction with the 20th International World Wide Web Conference in Hyderabad, India.* ACM Press, 2011.

[39] FastRandomForest. Re-implementation of the random forest classifier for the weka environment. `http://code.google.com/p/fast-random-forest/`.

[40] D. Fetterly and Z. Gyöngyi. Fifth international workshop on adversarial information retrieval on the web (AIRWeb 2009). 2009.

[41] D. Fogaras and B. Rácz. Scaling link-based similarity search. In *Proceedings of the 14th World Wide Web Conference (WWW)*, pages 641–650, Chiba, Japan, 2005.

[42] J. Fogarty, R. S. Baker, and S. E. Hudson. Case studies in the use of roc curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005*, GI '05, pages 129–136, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2005. Canadian Human-Computer Communications Society.

[43] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of statistics*, pages 337–374, 2000.

[44] G. Geng, X. Jin, and C. Wang. CASIA at WSC2008. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[45] X.-C. Z. Guang-Gang Geng, Xiao-Bo Jin and D. Zhang. Evaluating web content quality via multi-scale features. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[46] Z. Gyöngyi and H. Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, October 2005.

[47] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, 2005.

[48] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 576–587, Toronto, Canada, 2004.

[49] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[50] A. Hotho, D. Benz, R. Jäschke, and B. Krause, editors. *Proceedings of the ECML/PKDD Discovery Challenge*. 2008.

[51] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 538–543, 2002.

[52] Y. joo Chung, M. Toyoda, and M. Kitsuregawa. A study of web spam evolution using a time series of web snapshots. In *AIRWeb '09: Proceedings of the 5th international workshop on Adversarial information retrieval on the web*. ACM Press, 2009.

[53] C. Kohlschütter, P. A. Chirita, and W. Nejdl. Efficient parallel computation of pagerank, 2007.

[54] Z. Kou and W. W. Cohen. Stacked graphical models for efficient inference in markov random fields. In *SDM 07*, 2007.

[55] Y. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. Splog detection using content, time and link structures. In *2007 IEEE International Conference on Multimedia and Expo*, pages 2030–2033, 2007.

[56] T. Lynam, G. Cormack, and D. Cheriton. On-line spam filter fusion. *Proc. of the 29th international ACM SIGIR conference on Research and development in information retrieval*, pages 123–130, 2006.

[57] A. Niculescu-Mizil, C. Perlich, G. Swirszcz, V. Sindhwani, Y. Liu, P. Melville, D. Wang, J. Xiao, J. Hu, M. Singh, et al. Winning the KDD Cup Orange Challenge with Ensemble Selection. In *KDD Cup and Workshop in conjunction with KDD 2009*, 2009.

[58] V. Nikulin. Web-mining with wilcoxon-based feature selection, ensembling and multiple binary classifiers. In *Proceedings of the ECML/PKDD 2010 Discovery Challenge*, 2010.

[59] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 83–92, Edinburgh, Scotland, 2006.

[60] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *In Proceedings of SIGIR'94*, pages 232–241. Springer-Verlag, 1994.

[61] G. Shen, B. Gao, T. Liu, G. Feng, S. Song, and H. Li. Detecting link spam using temporal information. In *ICDM'06.*, pages 1049–1053, 2006.

[62] D. Siklósi, B. Daróczy, and A. Benczúr. Content-based trust and bias classification via biclustering. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 41–47. ACM, 2012.

[63] A. Singhal. Challenges in running a commercial search engine. In *IBM Search and Collaboration Seminar 2004*. IBM Haifa Labs, 2004.

[64] S. Webb, J. Caverlee, and C. Pu. Predicting web spam with HTTP session information. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 339–348. ACM, 2008.

[65] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.

[66] B. Wu, V. Goel, and B. D. Davison. Topical TrustRank: Using topicality to combat web spam. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, Edinburgh, Scotland, 2006.

[67] B. Zhou, J. Pei, and Z. Tang. A spamicity approach to web spam detection. In *Proceedings of the 2008 SIAM International Conference on Data Mining (SDM'08)*, pages 277–288. Citeseer, 2008.

Real-time streaming mobility analytics

András Garzó*, András A. Benczúr*, Csaba István Sidló*, Daniel Tahara†, Erik Francis Wyatt‡
* *Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA SZTAKI)*
Faculty of Informatics, University of Debrecen and Eötvös University, Budapest
Email: {garzo, benczur, scsi}@ilab.sztaki.hu
† *Yale University*
Email: daniel.tahara@yale.edu
‡ *St. Olaf College*
Email: wyatte@stolaf.edu

***Abstract*—Location prediction over mobility traces may find applications in navigation, traffic optimization, city planning and smart cities. Due to the scale of the mobility in a metropolis, real time processing is one of the major Big Data challenges.**

In this paper we deploy distributed streaming algorithms and infrastructures to process large scale mobility data for fast reaction time prediction. We evaluate our methods on a data set derived from the Orange D4D Challenge data representing sample traces of Ivory Coast mobile phone users. Our results open the possibility for efficient real time mobility predictions of even large metropolitan areas.

***Keywords*-Mobility, Big Data, Data Mining, Visualization, Distributed Algorithms, Streaming Data**

I. Introduction

Intelligent Transportation is at the center of worldwide transport policies intending to consolidate efficient and sustainable transport systems and associated infrastructures. The belief is that smart systems can receive, manage and provide valuable information that will allow transport users and operators to deal with a vast variety of traffic situations: congestion, safety, tolling, navigation support, law enforcement, as well as environmental sustainability.

Real time traffic prediction, as opposed to offline city planning, requires processing the incoming data stream without first storing, cleaning and organizing it in any sense. Scalability and low latency are crucial factors to enable any future technology to deal with mobility traces. This situation pushes towards new algorithms (typically, approximate or distributed) and new computational frameworks (e.g., MapReduce, NoSQL and streaming data). In this paper, we show that location prediction algorithms can be implemented in a distributed streaming environment, and remarkably high throughput can be achieved with low latency using a properly designed streaming architecture.

We use the D4D Challenge Data Set[1] for our experiments. In our research the emphasis is on the algorithmic and software scalability of the prediction method. Although there exist publications with similar goals, even recent results [1] consider data sets of similar or smaller size compared to D4D. Furthermore, we multiplied the original data set to meet the requirements of a metropolitan area of several million people using mobile devices all day.

The rest of this paper is organized as follows. Section II is devoted to describing the streaming data processing software architecture. Section III shows how distributed mobility data stream processing can be implemented in this architecture using Storm. Section IV describes the D4D data set used for our experiments. In Section V we describe the elements of the modeling and prediction framework. In Section VI we give our results, both in terms of accuracy and scalability. Finally related results are summarized in Section VII.

II. Streaming Architecture

Figure 1 depicts the layered architecture of our proposed general mobility data processing solution. The system enables easy model implementation while relying on the scalability, low latency and fault tolerance of the underlying distributed data processing software.

Distributed stream processing frameworks have not yet reached the same level of maturity such as batch processing ones based on MapReduce and key-value stores. Certain mature frameworks such as Hadoop [18] reach the required level of scalability, but cannot provide mechanisms for streaming input and real time response. As it is yet unclear which programming model of distributed stream processing will reach the same maturity and acceptance, we build an abstract stream processing layer capable of using some of the competing alternatives. We indicated Storm and S4 in Figure 1 as the most promising ones.

Stream processing frameworks cannot directly guarantee to store history information as their processing modules may restart from scratch after failures. For example if a machine crashes that stores information on part of the users in memory, these users will be repartitioned to other machines with empty history by default. To ensure that the history is preserved over the processing nodes even in case of failures, we build a generic persistence module (bottom right side of Figure 1). We store information on users and cell towers needed for modeling in distributed key-value stores. The

The publication was supported by grant OTKA NK 105645, the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956), and the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project is supported by the European Union, co-financed by the European Social Fund. D. Tahara and E Wyatt participated in the project during their study abroad semester at AIT-Budapest http://www.ait-budapest.com/.

[1] http://www.d4d.orange.com/home

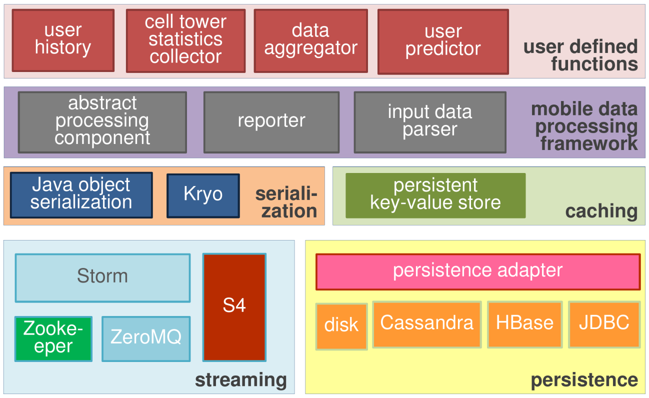


Figure 1. Layers of our mobility prediction architecture: the streaming framework (bottom left), persistence components (bottom right), and the custom analytics (top).

visualization dashboard also gets the required data through the storage adapter.

We defined a modular architecture for caching and serialization. Since near real time processing is very important, we deploy Cassandra [11] due to its high throughput writing capabilities and memcached [7] for its efficiency.

The mobility data processing layer (second from top in Figure 1) provides domain-specific primitives. For example, parsers, abstract data records and processing components for call detail records (CDRs) and other types of input are defined here. Built on these primitives, on the top of Figure 1, user defined functions implement history collection and location prediction. On the top level of our architecture, the implementation details of distributed processing, persistence, caching and serialization are hidden from the programmer to enable agile and straightforward model implementation.

III. Distributed Location Prediction Implementation

Our demonstration is based on Storm, a scalable, robust and fault tolerant, open source real time data stream processing framework developed and used by Twitter and many others [12]. Key to a practical application, Storm takes all the burden of scalability and fault tolerance.

We implement the required processing elements using the predefined abstract components of the Storm API: *spouts* responsible for creating input and *bolts* containing the processing step. Storm can distribute and balance multiple instances of spouts and bolts in the cluster automatically. Bolts can be connected to each other in acyclic processing graph by data packet streams as seen in Fig. 3.

Raw mobility data is read into the memory by an external application and is put into a lightweight message queuing system called Kestrel [4]: Storm spouts get their data from this buffer.

Key to our application is that partitioning can be controlled. As seen in Fig. 2, we may split incoming records both by user and by cell tower. Hence we may define

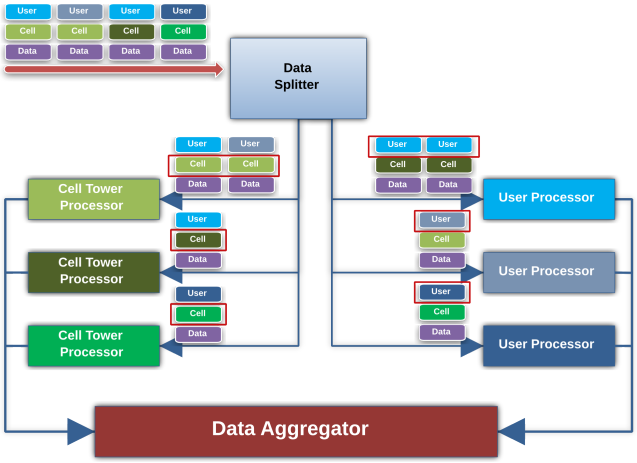


Figure 2. Sample input data partitioning in the streaming framework. Input records consist of tuples of user, cell tower and time stamp and may get partitioned both based on user and cell ID. User and cell based models may get merged through the data aggregator element of the streaming framework.

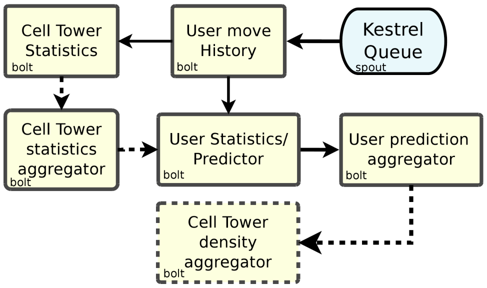

Cell Tower Statistics
bolt
User move History
bolt
Kestrel Queue
spout
Cell Tower statistics aggregator
bolt
User Statistics/ Predictor
bolt
User prediction aggregator
bolt
Cell Tower density aggregator
bolt

Figure 3. System block diagram of the Storm streaming components. Regular arrows show normal while dashed show the low frequency periodic special "tick" packets.

processing components both on the user and on the cell tower base. Finally user and cell tower models can be merged by using data aggregators.

We define two types of data flow, as seen in Fig. 3:

- One regular packet starts from the single spout for each input record that spreads along the thick edges in Fig. 3.
- Periodic aggregation updates move model information along the dashed edges initiated by the special built-in Storm *tick* packets.

We describe our algorithms by specifying the type of data moving along different edges of Fig. 3 and describing the algorithms implemented within each node of the Figure, the bolts that correspond to the steps described in Section V.

- The spout element emits tuples (u, a, t) of user, cell and time stamp.
- User history elements send sequences of (a, t) tuples of the past steps both to the last cell statistics bolt for recording the new user location and to the previous cell for counting frequencies through the cell.
- User history elements send trees rooted at the current

location (a, t) weighted with transition probabilities.

- Cell statistics elements periodically submit the frequent patterns to a single cell statistics aggregator bolt.
- The cell statistics aggregator bolt periodically refreshes the cell frequent patterns to all user statistics predictors.
- User statistics predictors emit the aggregated future history of the user in a form of rooted trees. This element is used in the current experiment to measure the accuracy of the user location prediction.
- User prediction aggregator periodically emits the predicted density of all cells seen in the prediction of the given user for aggregation by the single cell density aggregator element. In the current experiment this element measures the accuracy of the cell density prediction.

IV. The D4D Data Set

We used the Fine Resolution Mobility Trace Data Set (SET2) of the D4D Challenge [2], containing trajectories of randomly sampled individuals over two week periods. Trajectories have the granularity of cell tower locations. Only locations associated with voice, text or data traffic events are provided.

The SET2 data contains 50 million events. In a day, a large metropolis is expected to generate records two to three orders of magnitude more, especially if all locations related to all communication with the base stations is provided. Our target is to process events in the order of 100,000 in a second corresponding to several million people, each generating an event in every minute.

The fine-grained D4D data set is sparse to protect privacy. To reach the targeted amounts of data we merged the two week chunks of user location samples and considered the resulting data as if it is coming from a single two weeks period. The resulting weekly aggregated traffic volume is shown in Fig. 5, indicating that considering the time of the day only may be a good approximation for user motion.

The fact that only two-week user histories are available in the data set poses certain limitations for our experiments, however provides realistic distributions for scalability testing. In the data set the median users only visits three locations, and the mean only visits six. The median user generates 46 events, out of which changes location thirteen times. Most calls are in the same location as the previous call as seen in Fig. 4. We can achieve near 70% accuracy by always predicting the user's last location. In addition, we should only ever predict locations that a user has visited before and since we cannot see a smooth path for how a user moves over time, for now we ignore the physical layout of the antennas and treat location prediction as a classification problem.

V. Models for Location Prediction

We give sample models to predict user movement and traffic congestion. We produce a simple yet realistic framework for location prediction sufficiently complex for scalability testing. We predict sequences and evaluate always for the next known location after the predefined prediction period.

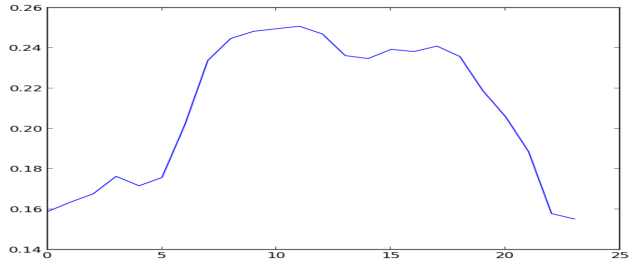


Figure 4. Fraction of calls that are at a different antenna than the previous call for that user (y axis) by time of day (x axis). We can see morning and evening rush hour, and people move less at night. Even at the peak of morning rush hour, more then three quarters of all calls are from the same location as before.

The main component of our model is based on learning the patterns of individual user motion histories. Our main assumption is that for most users, their location follows a daily regular schedule, i.e. they leave to work and return home at approximately the same time of the day over roughly the same route. This assumption is confirmed for other data sets e.g. in [9]. We consider typical locations and two-step frequent patterns. For each user, we generate the following set of features:

- Time of the day;
- Time elapsed since the previous location;
- Ratio of events from frequently used antennas;
- Typical locations at the time of the day and distance from previous location;
- Typical length of stay in one place in general and depending on time of the day.

The last two classes of features are implemented by nearest neighbor search among blocks of events consisting of subsequent events from the same location. Distance is calculated by taking the time of the day, the duration of stay at the same location, the geographical distance and the equality of the present and the past antennas. We compute the nearest two neighbors under four different selection of these attributes:

(A) Time of the day only;
(B) Time of the day and equality of the past cell tower;
(C) Duration of stay and distance from the previous cell tower;
(D) Time of the day, duration of stay and distance from the previous cell tower.

Based on the above feature set, we use decision trees for modeling. First we predict whether the user changed or remained in the same location. For location prediction we have no information other than user past locations and frequent paths through user most recent locations. Hence we train classifiers separate for each of the following possibilities for the next location:

- Same as previous location;
- Most frequent (second most frequent) location;

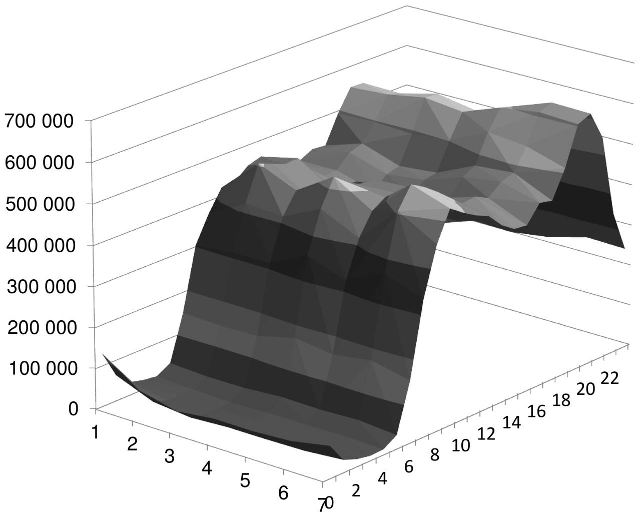


Figure 5. Volume of traffic (vertical axis) as the function of the day of the week (1–7) and hour (0–23) over the horizontal axes.

- One of the nearest neighbor locations;
- Next step over frequent paths—here longer paths could also be computed, e.g. by streaming FP-trees [8].

We consider the first week as training period: for each user event, we compute all the above features based on the user history before the event. We give a set of binary labels to the event to mark whether the user stayed in the previous location or moved to one of the potential new locations. As additional features, we also compute the physical distance of the last location to each of the potential new ones.

In our implementation, the modeling steps correspond to the Storm bolts of Fig. 3 as follows. User features are computed based on past history in the `user move history` bolt. In order to compute frequent paths, the `cell tower statistics` bolt receives the last few user steps from the `user move history` bolt. Frequent paths need only be periodically updated and this is done in the `cell tower statistics` bolt that feeds the `user statistics/predictor` bolt with updates. This bolt is capable of implementing the pre-trained decision tree model.

VI. Experiments

In this section we describe our measurements for speed, scalability and quality. To emphasize scalability in the number of threads and machines, we ran our experiments over a Storm 0.9.0-wip4 cluster of 35 old dual core Pentium-D 3.0GHz machines with 4GB RAM each.

The spouts emit as many records as the Storm framework is able to process. We partitioned the data for the spouts and for each spout, we loaded its data set into memory while initializing the topology. We iterated over the data infinitely, i.e. the same user moves were emitted repeatedly.

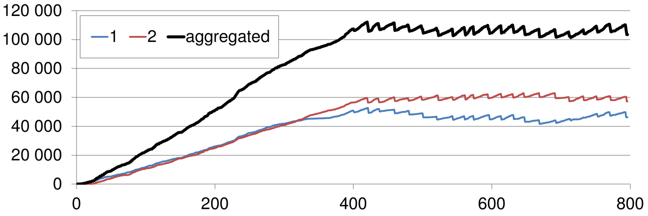


Figure 6. Number of records emitted by two spouts per 13 minutes (vertical axis, records per second) after initializing the topology (with seconds on the horizontal axis). Red and blue lines indicate throughput of two spouts and the black bold line is the aggregated speed.

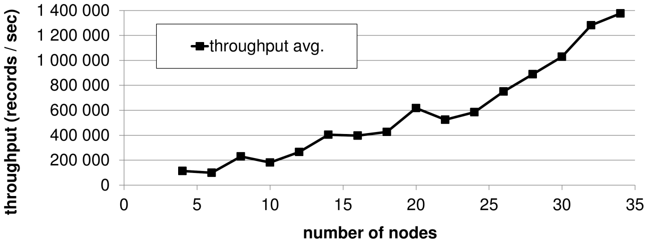


Figure 7. Throughput (number of records per second) as the function of the number of servers in the Storm cluster, with five input spouts residing at five different servers.

A. Scalability and Latency

To test scalability of location prediction we test how the throughput (the number of events processed per second) changes when new nodes are added to the cluster. To avoid misleading figures due to caching, we ran the system for 10 minutes before starting to measure the predictor element processing rate. Figure 6 shows how system throughput normalizes after initialization.

Figure 7 depicts throughput speed. Near linear scalability can be observed in the number of servers and threads: We may reach rates of a few 100,000 records in a second, which is well beyond the desired threshold.

The average latency of the system was low, processing an input record took about 1023 ms. We did observe larger values when initializing the system, but this value remained relatively constant when adding or removing nodes.

B. Fault Tolerance

When a node fails, a new node is initialized with the stored states of the affected processing components. According to the guarantees of Storm, the lost packets are also processed again. Figure 8 shows how node failures affect overall performance. We can observe rapid recoveries, despite of the large number of failing nodes, the overall performance remains predictable.

C. Accuracy

Next we evaluate the accuracy of our location prediction methods by giving F-measure (averaged for the positive and negative classes) and AUC values. We predict the location of active users for at least 15 minutes in the future. We

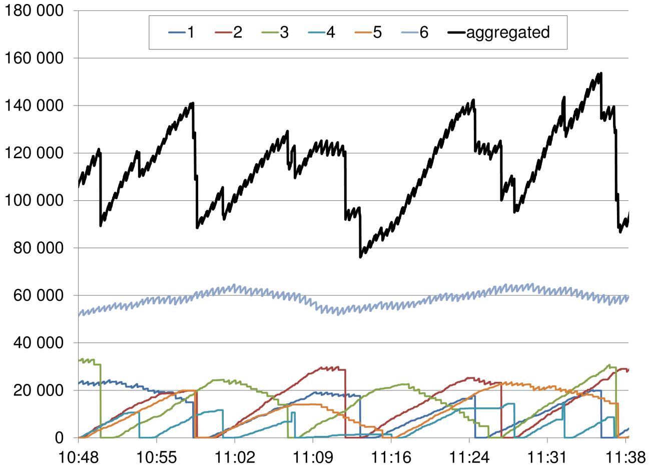


Figure 8. Throughput (number of records per second) as the function of the time passed (absolute times), with six nodes, each with a spout. One node works continuously, while the others occasionally stop.

	All		Active	
	F	AUC	F	AUC
next	0.888	0.914	0.888	0.886
15 mins	0.859	0.905	0.819	0.896
1 hour	0.861	0.909	0.815	0.890

Table I
ACCURACY OF THE PREDICTION WHETHER A GIVEN USER MAKES THE NEXT CALL FROM A NEW LOCATION.

consider a user active if he or she has at least 1000 events during the two-week period. There are 1126 such users in the data set. We use the first week of data for all users for training and evaluate over the second week.

The prediction for users staying in place is given in Table I. Here we observe that the prediction quality is very high and slightly decays as we look farther ahead in the future. The decision tree has 37 nodes using the following sample of attributes in approximate order of tree depth:

- Previous location equal to most frequent user cell;
- Fraction of the last and the most frequent cells in the user history so far;
- Geographical distance and duration of stay at nearest neighbor (D) and other nearest neighbors in case of the active users;
- Elapsed time since arrival to the last location.

The prediction for the next user location is evaluated for active users for a minimum of 15 minutes in the future. As seen in Table II, we perform binary classification tasks for ten different types of likely next locations for the user. Note that some of these locations may coincide or not exist, hence no multi-class classification is possible. Based on the measured accuracy of the methods and the likelihood assigned by the decision tree, it is easy to merge the binary predictions into a prediction of a single location.

The decision for the most frequent continuation of the last two cell locations is weak, however misclassification is imbalanced: we almost never misclassify users who do not follow the frequent path. Here the decision tree is surprisingly small, it has only five leaves. The first decision splits whether the user stayed for more or less than one day at the same location. Subsequent decisions use the fraction of the previous cell among all cells of the user and the distance of the last step taken.

	F	AUC
same as previous	0.862	0.896
most frequent 3-step trajectory	0.372	0.623
nearest neigbor (A)	0.637	0.686
second nearest neigbor (A)	0.633	0.633
nearest neigbor (B)	0.710	0.699
second nearest neigbor (B)	0.700	0.695
nearest neigbor (C)	0.708	0.705
second nearest neigbor (C)	0.698	0.695
nearest neigbor (D)	0.553	0.686
second nearest neigbor (D)	0.672	0.669

Table II
ACCURACY OF THE PREDICTION FOR DIFFERENT TYPES OF NEW LOCATIONS AS DESCRIBED IN SECTION V.

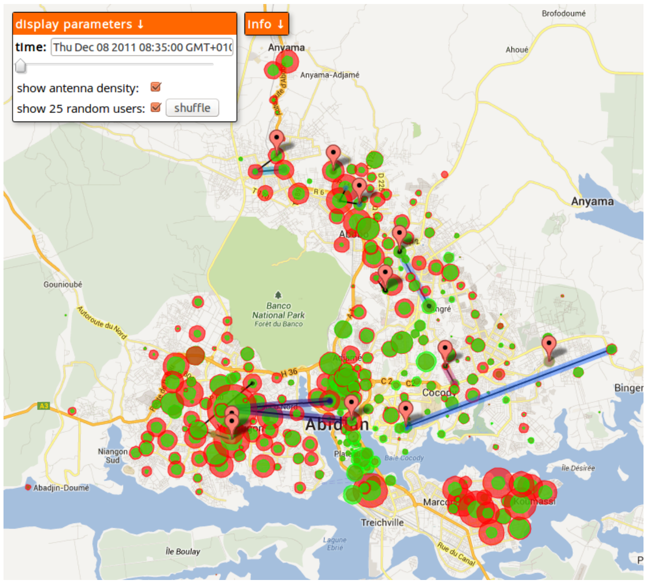

display parameters ↓
time: Thu Dec 08 2011 08:35:00 GMT+010
show antenna density: ☑
show 25 random users: ☑ shuffle
Info ↓

Figure 9. Visualization of the cell traffic prediction (red circles show actual sizes while green is the prediction), with a sample of individual movement predictions (black lines are real, colored lines are predicted moves).

For nearest neighbors, the decision trees mostly choose based on physical distance. This is the main reason why we see very similar measures for the last eight classification problems. In addition, some features to determine whether the user stays in place are also used but in general the decision trees are much smaller than for staying in place.

We developed a visualization demo application to demonstrate the use of individual trajectory predictions: Fig. 9 shows the aggregated predicted and real cell density as well as the predicted and real trajectories of random users.

VII. Related Results

The idea of using mobility traces for traffic analysis is not new. Early papers [14] list several potential applications, including traffic services, navigation aids and urban system mapping. In [14] a case study of Milan, while in [1] of New York City suburbs are presented.

Mobility, City Planning and Smart City are considered by many major companies. IBM released redguides [10], [15] describing among others their IBM Traffic Prediction Tool [15]. Unfortunately little is known about the technology and the scalability properties of existing proprietary software.

Several recent results [9], [16], [3], [17], [19] analyze and model the mobility patterns of people. In our results we rely on their findings, e.g. the predictability of user traces.

We are aware of no prior results that address algorithmic and software issues of the streaming data source. Mobility data naturally requires stream processing frameworks [12], [13], [5]. A wide range of stream processing solutions are available: For example, major social media companies have all developed their software tools [6].

VIII. Conclusions and Future Work

In this preliminary experiment we demonstrated the applicability of data streaming frameworks for processing mass mobility streams: Low latency and high throughput values enable building real-time applications based on motion prediction. Given more detailed data, our framework is suitable for detecting flock (motion of groups), deviation of real track from expected (map) or permitted (restricted areas) tracks. Our results open the ground for advanced experimentation regarding the quality of large scale mobility prediction suitable for example for real time traffic prediction.

Acknowledgment

Among several colleagues who draw our attention to the scalability issues of mobility data and inspired our research, we are especially grateful, in alphabetic order, to Tamas Bodis (Vodafone Hungary), Csaba Fekete (NSN Hungary), Zalan Heszberger (Ericsson Hungary and TU Budapest), Attila Medvig (NNG) and Janos Tapolcai (TU Budapest).

References

[1] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. *Pervasive Computing, IEEE*, 10(4):18–26, 2011.

[2] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the D4D challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

[3] de Montjoye, Yves-Alexandre, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Nature Sci. Rep.*, 2013.

[4] E. D. Dolan, R. Fourer, J.-P. Goux, T. S. Munson, and J. Sarich. Kestrel: An interface from optimization modeling systems to the neos server. *INFORMS Journal on Computing*, 20(4):525–538, 2008.

[5] M. Dusi, N. d'Heureuse, F. Huici, A. di Pietro, N. Bonelli, G. Bianchi, B. Trammell, and S. Niccolini. Blockmon: Flexible and high-performance big data stream analytics platform and its use cases. *NEC Technical Journal*, 7(2):103, 2012.

[6] D. Eyers, T. Freudenreich, A. Margara, S. Frischbier, P. Pietzuch, and P. Eugster. Living in the present: on-the-fly information processing in scalable web architectures. In *Proceedings of the 2nd International Workshop on Cloud Computing Platforms*, page 6. ACM, 2012.

[7] B. Fitzpatrick. Distributed caching with memcached. *Linux journal*, 2004(124):5, 2004.

[8] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu. Mining frequent patterns in data streams at multiple time granularities. *Next generation data mining*, 212:191–212, 2003.

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[10] M. Kehoe, M. Cosgrove, S. Gennaro, C. Harrison, W. Harthoorn, J. Hogan, J. Meegan, P. Nesbitt, and C. Peters. Smarter cities series: A foundation for understanding ibm smarter cities. *Redguides for Business Leaders, IBM*, 2011.

[11] A. Lakshman and P. Malik. Cassandra: A structured storage system on a p2p network. In *Proceedings of the twenty-first annual symposium on Parallelism in algorithms and architectures*, pages 47–47. ACM, 2009.

[12] J. Leibiusky, G. Eisbruch, and D. Simonassi. *Getting Started With Storm*. Oreilly & Associates Incorporated, 2012.

[13] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed stream computing platform. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 170–177. IEEE, 2010.

[14] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, 33(5):727, 2006.

[15] S. Schaefer, C. Harrison, N. Lamba, and V. Srikanth. Smarter cities series: Understanding the ibm approach to traffic management. *Redguides for Business Leaders, IBM*, 2011.

[16] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[17] L. A. Tang, Y. Zheng, J. Yuan, J. Han, A. Leung, C.-C. Hung, and W.-C. Peng. On discovery of traveling companions from streaming trajectories. In *ICDE*, pages 186–197, 2012.

[18] T. White. *Hadoop: The Definitive Guide*. Yahoo Press, 2010.

[19] K. Zheng, Y. Zheng, N. J. Yuan, and S. Shang. On discovery of gathering patterns from trajectories. In *ICDE*, pages 242–253, 2013.