Network Analysis for Wikipedia

F. Bellomi and R. Bonato

Dipartimento di Informatica, Università di Verona

Cà Vignal, 2 Strada Le Grazie, 15 I-37134 Verona (Italy)

francesco.bellomi@gmail.com, robonato@tiscali.it

Abstract

Network analysis is a quantitative methodology for studying properties related to connectivity and distances in graphs, with diverse applications like citation indexing and information retrieval on the Web. The hyperlinked structure of Wikipedia and the ongoing, incremental editing process behind it make it an interesting and unexplored target domain for network analysis techniques.

In this paper we apply two relevance metrics, HITS and PageRank, to the whole set of English Wikipedia entries, in order to gain some preliminary insights on the macro-structure of the organization of the corpus, and on some cultural biases related to specific topics.

1 Introduction

Wikipedia is a free, open-content, online encyclopedia. There is virtually no restriction on the content of its articles, which range from classical encyclopedic topics on art, history and science, to more mundane ones like tv stars biographies, urban legends and breaking news. Its content is entirely provided by volunteers who are free to create, edit and revise any entry. The process is governed by Wikipedia's official neutral point of view (NPOV) policy, which requires that contributors work to avoid bias in writing articles. Contributors build upon each other's changes and flawed edits are repaired in a constantly on-going reviewing process. To date, Wikipedia features more than 600,000 articles in English, and more than 1.5 million entries as a multilingual resource.

Wikipedia is structured as an interconnected network of articles. Each article can hyperlink several other Wikipedia entries. It's up to the contributor to establish a hyperlink connection between a term that occurs in the article and the corresponding Wikipedia entry, provided that one exists; if the entry does not exist, it is possible to create a new empty "stub" entry, and link it.

Wikipedia is a growing repository of distributed and interconnected knowledge. Its hyperlinked structure and the ongoing, incremental editing process behind it make it very

peculiar domain to study the impact of Network Analysis techniques to a restricted, controlled corpus where resources are not generic web pages, but content relevant entries edited and constantly revised by a community of committed users.

We expect quantitative tools and methods from network analysis to provide a new, quantitatively founded perspective into this very peculiar artifact of human knowledge.

2 Network Analysis and Content Relevance Metrics

Network analysis is a branch of graph theory which aims at describing quantitative properties of networks of interconnected entities by means of mathematical tools. Any domain which can be described as a set of interconnected objects is a domain application for network analysis. Its methods and tools work on top of this abstraction, and as such they are totally indifferent to the nature and properties of the entities involved, be they train stops in a railway network, individuals of a given social group bound by kinship relationship, or hosts in a computer network. In particular, network analysis has recently provided successful algorithms to tackle some important problems connected to Internet search technologies.

A common search engine can very well return thousands of webpages as the answer to a single query. In order for the user to be able to quickly identify what answer best matches the query, results must be ranked according to a relevance criterion. In-depth content analysis of the results is neither effective nor efficient for a task which must be accomplished in fractions of seconds over tens of thousands of webpages. Network analysis provides content-independent effective metrics for relevance which exclusively rely on the analysis of hyperlink structure of results. In the following subsections we quickly sketch two of the most popular algorithms issued from network analysis to perform what is commonly known as *content relevance ranking*: Jon Kleinberg's HITS and Larry Page and Sergey Brin's PageRank algorithm.

2.1 The HITS algorithm

HITS (Hyperlink-Induced Topic Selection, see [1]) is an algorithm for ranking web pages related to a common topic according to their potential relevance. HITS exclusively relies on the hyperlink relations existing among the pages of a given domain (in the case of web-search engines, the whole Internet), which are used to define two mutually reinforcing weights called *hub* and *authority*. Intuitively, a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs. It can be shown that for each page these two mutually reinforcing measures converge to fixed points, which will be the hub and authority weights for the page. Authority is used to rank pages resulting from a given query (and thus potentially related to a given topic) in order of relevance. The intuitive idea behind the algorithm is that if pages i and j belong to a set of pages supposedly meaningful on a given topic, a link from page i towards page j gives "authority" to j on that topic. However, the authority provided by i to j is proportional

to its weight as a "hub", that is, its ability to point at many other good authorities.

We can view any collection V of hyperlinked pages as a directed graph $G = (V, E)$: the nodes correspond to the pages, and a directed edge $(p, q) \in E$ indicates the presence of a link from p to q. Let S be a set of web pages returned by a query on a search engine. So each page p has a non-negative *authority weight* $x^{\langle p \rangle}$ and a non-negative *hub weight* $y^{\langle p \rangle}$ associated. Weights are normalized so their squares sum to 1: $\sum_{p \in S} (x^{\langle p \rangle})^2 = 1$, and $\sum_{p \in S} (y^{\langle p \rangle})^2 = 1$.

The informal notion that we sketched at the beginning of the section can be expressed as a mutually reinforcing relationship between hubs and authorities. If p points to many pages with large $x-$values, then it should receive a large $y-$value; and if p is pointed to by many pages with large $y-$values, then it should receive a large $x-$value. Hubs and authorities are calculated via an iterative algorithm that maintains and updates numerical weights for each page. This is implemented by introducing two operations on the weights, which are denoted by $\mathcal{I}$ and $\mathcal{O}$. Given weights $x^{\langle p \rangle}$, $y^{\langle p \rangle}$, the $\mathcal{I}$ and $\mathcal{O}$ operations update the $x-$weights and $y-$weights, respectively, as follows.

$$\mathcal{I}: \quad x^{\langle p \rangle} \leftarrow \sum_{q:(q,p) \in E} y^{\langle p \rangle} \qquad\qquad \mathcal{O}: \quad y^{\langle p \rangle} \leftarrow \sum_{q:(q,p) \in E} x^{\langle p \rangle}$$

By means of operations $\mathcal{I}$ and $\mathcal{O}$, hubs and authorities get mutually reinforced. Jon Kleinberg proves in [1] that by applying the $\mathcal{I}$ and $\mathcal{O}$ operations in an alternating fashion, hub and authority weights eventually converge to an "equilibrium point", that is a fixed point is reached, which represents the actual authority and hub weight for each page.

2.2 Pagerank

PageRank algorithm (which lies at the core of the popular search engine Google, see [2]) relies on the vast link structure of the Web as an indicator of an individual page's relevance metric. In the words of its authors, Pagerank interprets a link from page A to page B as a "vote" by page A for page B. However, analogously to the HITS algorithm, votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

The formula uses a model of a random surfer who gets bored after several clicks and switches to a random page. The PageRank value of a page reflects the frequency of hits on that page by the random surfer. It can be understood as a Markov process in which the states are pages, and the transitions are all equally probable and are the links between pages. If a page has no links to another pages, it becomes a sink that will trap the random visitors forever. So the models assumes that if the random surfer arrives to a sink page, it picks another URL at random and continues surfing again. To be fair with pages that are not sinks, these random transitions are added to all nodes in the Web, with a residual probability of usually q=0.15, estimated from the frequency that an average surfer uses his or her browser's bookmark feature.

Informally stated, it can be said that the PageRank algorithm calculates the quality of a page p_i proportionally to the quality of the pages that contain in-links to it. Since

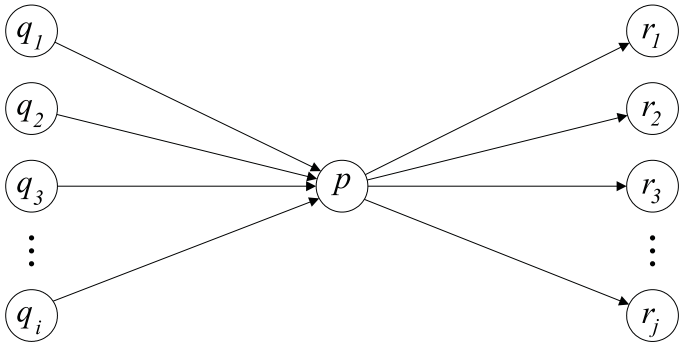

Figure 1: The PageRank score for a page p depends on the PageRank scores of referring pages, q_1 to q_i pointing to p. In the HITS algorithm, the authority score of a page p depends on the hub scores of referring pages, q_1 to q_i, and the hub score depends on the authority scores of the pages to which p is pointing, r_1 to r_j.

PageRank also considers the quality of the in-linking pages, the score of p_i is calculated recursively as follows:

$$\text{PageRank}(p_i) = \frac{q}{N} + (1-q) \sum_{p_j \in M(p_i)} \frac{\text{PageRank}(p_j)}{L(p_j)}$$

where $p_1, p_2, ..., p_N$ are the pages under consideration, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of links coming from page p_j, and N is the total number of pages. In can be proved that even this metrics converges quite efficiently after a limited number of iterations to a fixed point which represents the PageRank value for any page.

3 Performing network analysis on Wikipedia

In order to compute HITS and PageRank metrics on Wikipedia's entries, we take into consideration the link graph consisting of the HTML hyperlinks that are 1) contained in the entries' definition , and 2) which point to another Wikipedia entry. Both conditions are enforceable in a precise way, by exploiting some specific conventions of the content markup generated by MediaWiki, the software platform that powers Wikipedia.

Given the nature of HTML, only outbound links are explicitly stated, so inbound links are computed from the outbound links; this means that only inbound links from within the encyclopedia are taken into consideration, which is a desirable feature.

Outbound links to non-existent stub entries (that is, the links highlighted in red in Wikipedia's pages) are not taken into consideration: although they could theoretically be used within the analysis framework, as virtual nodes with no outbound links, we maintain that they does not represent real "content", and we think that they would provide a misleading contribution to both metrics: they draw out relevancy from the referrers without giving it back to any other referee.

We also applied some heuristics to merge the nodes related to duplicate entries that appear only with slight variations in their names, like different capitalization, of a different choice of punctuation.

We removed from the graph all the nodes related to entries with titles that start with "List of...". These are basically only collections of links, that we maintain to be a source of noise with respect to our goal of providing a measure of the relevancy of entries: an entry is cited in this kind of lists not because it is maintained by the author as relevant and appropriate to describe the meaning of a term, but only for the mere fact that it belongs to a collection of terms; and this collection itself it is not a proper "concept", since it does not even have a real name, other that "List of...".

We have developed a specialized web crawler in order to download the whole content of the English language Wikipedia. The crawler scans the "All pages" index section of Wikipedia, thus retrieving a complete list of the pages exposed on the web site; the "special pages" (that is, the pages not containing an entry) are then filtered out, by looking at specific patterns in the URLs, and then all the proper pages have been downloaded and stored locally, in order to perform the computation of the metrics. All the results discussed in this paper come from a "snapshot" of the Wikipedia corpus taken in a period of 34 hours, between the 16th and the 17th of April 2005.

4 Results

The aim of our simple experiment is twofold: to gain some understanding of the high level *structure* of Wikipedia, and to get some insights about its *content*, and in particular on its hidden cultural biases. Each user usually browse a (relatively) small set of entries during the normal usage of an encyclopedia; and such small sample is more representative of the enquiring user world view, rather that the whole encyclopedia. As a consequence, nobody is really able to have a mile-high fisheye view of Wikipedia's content; of course it is possible to perform some basic statistic analysis (like counting the entries, or measuring the rate of growth) but these are purely syntactic measures. We maintain that network analysis offers a simple way to have some more "semantic" measures, since it formally analyzes an intrinsically semantic human-generated type of content: the use of terms to define other terms.

A first important result is that Wikipedia's internal references form a single connected graph: there are no separated "islands" of entries with no inbound or outbound links to the rest of the corpus; given any entry, it is possible to reach any other entry, following a path of undirected reference links.

We adopt the following methodology: for each relevancy metric, we try to identify what classes of concepts are marked as most relevant in the global ranking. This should provide some hints on the general structure of Wikipedia. Then, we isolate some specific classes of entries (namely Countries and Cities, Historical Events, People and Common Nouns) and extract a topic-specific ranking, in order to try to identify specific cultural biases.

4.1 HITS

The HITS algorithm computes two distinct metrics for each node: *authority* and *hubness*. We use only the first, since it is proven to be the most effective as a relevancy metric; in fact hubness is often considered only instrumental to the computation of authority.

Here is the list of the 300 most relevant entries, according to the HITS authority metric:

1. United States, 2. France, 3. United Kingdom, 4. Germany, 5. Canada, 6. England, 7. Australia, 8. Japan, 9. Italy, 10. World War II, 11. Europe, 12. Spain, 13. Russia, 14. India, 15. Netherlands, 16. Sweden, 17. London, 18. Poland, 19. Mexico, 20. Belgium, 21. Soviet Union, 22. Switzerland, 23. Austria, 24. China, 25. New Zealand, 26. Greece, 27. Norway, 28. 1980s, 29. Paris, 30. Denmark, 31. World War I, 32. New York City, 33. California, 34. Portugal, 35. Turkey, 36. 1970s, 37. South Africa, 38. 1960s, 39. New York, 40. Finland, 41. Brazil, 42. 1990s, 43. Ireland, 44. Hungary, 45. Israel, 46. Egypt, 47. 19th century, 48. Television, 49. English language, 50. North America, 51. 20th century, 52. Africa, 53. Cuba, 54. Argentina, 55. Scotland, 56. Romania, 57. Ukraine, 58. People's Republic of China, 59. Philippines, 60. Asia, 61. 1950s, 62. Scientific classification, 63. Czech Republic, 64. Pakistan, 65. Animal, 66. Iraq, 67. United Nations, 68. Iran, 69. Bulgaria, 70. Population, 71. European Union, 72. South Korea, 73. Jew, 74. Indonesia, 75. Croatia, 76. French language, 77. Lithuania, 78. Cyprus, 79. Chile, 80. Thailand, 81. Slovenia, 82. Square kilometre, 83. Republic of Ireland, 84. Iceland, 85. Rome, 86. Singapore, 87. Slovakia, 88. Afghanistan, 89. Hong Kong, 90. Estonia, 91. Luxembourg, 92. Texas, 93. 1930s, 94. Malaysia, 95. President of the United States, 96. Latvia, 97. Chordate, 98. Puerto Rico, 99. Colombia, 100. Venezuela, 101. Florida, 102. Population density, 103. Jamaica, 104. Berlin, 105. Area, 106. Malta, 107. Hawaii, 108. Morocco, 109. Belarus, 110. Public domain, 111. Christianity, 112. Syria, 113. United States Navy, 114. Algeria, 115. Capital, 116. South America, 117. Albania, 118. Actor, 119. Peru, 120. Vietnam, 121. New Jersey, 122. Islam, 123. Nigeria, 124. Virginia, 125. United States Republican Party, 126. Saudi Arabia, 127. United States Democratic Party, 128. German language, 129. BBC, 130. Massachusetts, 131. Kenya, 132. 1920s, 133. George W. Bush, 134. Bird, 135. Film, 136. Pennsylvania, 137. Lebanon, 138. 18th century, 139. Quebec, 140. Britain, 141. Ontario, 142. United States House of Representatives, 143. Alaska, 144. Wales, 145. Latin, 146. Great Britain, 147. Moscow, 148. Ethiopia, 149. Cold War, 150. Middle East, 151. 1911 Encyclopdia Britannica, 152. Currency, 153. Panama, 154. Time zone, 155. Republic of China, 156. Atlantic Ocean, 157. Christian, 158. British, 159. Taiwan, 160. Tunisia, 161. Sri Lanka, 162. Libya, 163. Uruguay, 164. Illinois, 165. NATO, 166. United States Senate, 167. Spanish language, 168. American Civil War, 169. Washington, 170. Dominican Republic, 171. Pacific Ocean, 172. Jordan, 173. Ohio, 174. Haiti, 175. North Korea, 176.

Azerbaijan, 177. Serbia and Montenegro, 178. Nicaragua, 179. Radio, 180. Kuwait, 181. Native American, 182. Guatemala, 183. Sudan, 184. Bangladesh, 185. Oregon, 186. Bolivia, 187. Czechoslovakia, 188. Ecuador, 189. 1940s, 190. Trinidad and Tobago, 191. Vienna, 192. Automobile, 193. Ghana, 194. Uganda, 195. Barbados, 196. Monaco, 197. United States Congress, 198. Bahamas, 199. Arizona, 200. Zimbabwe, 201. Density, 202. National anthem, 203. Agriculture, 204. Tanzania, 205. Tokyo, 206. Michigan, 207. Costa Rica, 208. Internet, 209. Bosnia and Herzegovina, 210. Vietnam War, 211. Motto, 212. 17th century, 213. Georgia (country), 214. Dpartement, 215. Korea, 216. Louisiana, 217. Liechtenstein, 218. Colorado, 219. Athens, 220. Top-level domain, 221. Mozambique, 222. Muslim, 223. Armenia, 224. Minnesota, 225. Greek language, 226. 16th century, 227. Law, 228. Angola, 229. Caribbean, 230. North Carolina, 231. Namibia, 232. Roman Empire, 233. Georgia (U.S. state), 234. Cameroon, 235. Gibraltar, 236. Papua New Guinea, 237. Cambodia, 238. Senegal, 239. Jazz, 240. Religion, 241. El Salvador, 242. June, 243. 2004 Summer Olympics, 244. May, 245. Communism, 246. Maryland, 247. Adolf Hitler, 248. Sydney, 249. Guam, 250. Connecticut, 251. August, 252. Mauritius, 253. Economics, 254. Ottoman Empire, 255. Parliament, 256. July, 257. Moldova, 258. Missouri, 259. Uzbekistan, 260. United Arab Emirates, 261. Maine, 262. Zambia, 263. Utah, 264. Government, 265. Belize, 266. Bill Clinton, 267. Grenada, 268. Olympic Games, 269. Latin America, 270. Football (soccer), 271. Species, 272. British Columbia, 273. Tennessee, 274. Madagascar, 275. Honduras, 276. United States Army, 277. Nazi Germany, 278. U.S. state, 279. Guyana, 280. South Carolina, 281. Music, 282. Sierra Leone, 283. Boxing, 284. Independence, 285. Prussia, 286. Andorra, 287. Earth, 288. Basketball, 289. March, 290. Munich, 291. Kentucky, 292. Northern Ireland, 293. Qatar, 294. Democracy, 295. Official language, 296. San Marino, 297. Fiji, 298. French Revolution, 299. God, 300. Philosophy

It is apparent that the HITS authority metric reveal space (in the form of political geographical denominations) and time (in the form of both time spans and landmark historical events) to be the main organizing categories for Wikipedia. On a lesser extent, famous people (the first one mentioned is George W. Bush), common words such as Television, Animal, ethnical groups (the first one mentioned is Jew), political and social institutions and organizations, and abstract nouns, such as Music, Philosophy, Religion, and so on. Proper nouns are the vast majority, but this is somehow expected within an encyclopedia.

4.2 PageRank

Here is the list of the 300 most relevant entries, according to the PageRank metric:

1. united states, 2. christianity, 3. roman catholic church, 4. 2004, 5. eastern orthodox church, 6. jesus, 7. greek language, 8. russia, 9. bishop, 10. rome, 11.

canada, 12. europe, 13. pope, 14. japan, 15. latin, 16. god, 17. constantinople, 18. bible, 19. judaism, 20. oriental orthodoxy, 21. new testament, 22. united kingdom, 23. jerusalem, 24. eastern rite, 25. france, 26. world war ii, 27. greece, 28. protestantism, 29. priest, 30. egypt, 31. church, 32. protestant, 33. eastern orthodoxy, 34. islam, 35. roman empire, 36. apostle, 37. wikisource, 38. china, 39. full communion, 40. old testament, 41. antioch, 42. paul of tarsus, 43. patriarch of constantinople, 44. easter, 45. autocephaly, 46. trinity, 47. eucharist, 48. canon law, 49. holy spirit, 50. saint peter, 51. clergy, 52. baptist, 53. salvation, 54. faith, 55. christmas, 56. byzantine empire, 57. abraham, 58. germany, 59. patriarch, 60. anglicanism, 61. messiah, 62. heresy, 63. monastery, 64. eastern christianity, 65. ecumenical council, 66. apocrypha, 67. altar, 68. deacon, 69. lutheranism, 70. methodism, 71. east-west schism, 72. incarnation, 73. history of christianity, 74. pentecost, 75. laity, 76. ten commandments, 77. lent, 78. holy see, 79. evangelicalism, 80. cathedral, 81. religious minister, 82. liberal christianity, 83. christian ecumenism, 84. elder "religious", 85. bethlehem, 86. divine grace, 87. syria, 88. religious society of friends, 89. christian theology, 90. nazareth, 91. western christianity, 92. charismatic, 93. basilica, 94. 5th century, 95. council of chalcedon, 96. 1970, 97. bulgaria, 98. fundamentalist christianity, 99. liturgical year, 100. pastor, 101. god the father, 102. pentecostalism, 103. mary magdalene, 104. reformed churches, 105. holy day of obligation, 106. restorationism, 107. christian liturgy, 108. fall "religion", 109. 2003, 110. new testament view on jesus111. christian eschatology, 112. creation according to genesis, 113. chapel, 114. hillel the elder, 115. serbian orthodox church, 116. modernist christianity, 117. beatitudes, 118. russian orthodox church, 119. prayer in christianity, 120. pulpit, 121. christian philosophy, 122. christian worldview, 123. steeple 124. england, 125. 1920, 126. middle east, 127. alexandria, 128. roman catholicism, 129. moscow, 130. gregorian calendar, 131. synod, 132. primate "religion", 133. 7th century, 134. september 11, 135. schism, 136. coptic christianity, 137. orthodox church of constantinople, 138. 451, 139. augustine of hippo, 140. italy, 141. eastern europe, 142. icon, 143. arab, 144. 2005, 145. monk, 146. reformation, 147. antiochian orthodox church, 148. nestorianism, 149. apostles, 150. orthodox, 151. one holy catholic and apostolic church, 152. christ, 153. lutheran, 154. dogma, 155. assyrian church of the east, 156. apostolic succession, 157. tsar, 158. maronite, 159. aristotle, 160. syrian catholic church, 161. monophysite, 162. western world, 163. septuagint, 164. orthodox church of alexandria, 165. latin language, 166. 1453, 167. roman emperor, 168. orthodox church of jerusalem, 169. 1204, 170. monasticism, 171. 1054, 172. fall of constantinople, 173. orthodox church in america, 174. english language, 175. 2002, 176. holy orders, 177. tertullian, 178. slavic peoples, 179. divine liturgy, 180. fourth crusade, 181. constantine i of the roman empire, 182. papal infallibility, 183. russians, 184. centuries, 185. crucifixion, 186. pope paul vi, 187. ukrainians, 188. romanian orthodox church, 189. 2001, 190. russian orthodox

church outside russia, 191. cyril of alexandria, 192. bulgarian orthodox church, 193. iconoclasm, 194. second council of nicaea, 195. 19th century, 196. clerical celibacy, 197. filioque clause, 198. syro-malabar catholic church, 199. syriac orthodox church, 200. 787, 201. syro-malankara catholic church, 202. great lent, 203. slavic languages, 204. eastern orthodox church organization, 205. iconostasis, 206. ussr, 207. saint methodius, 208. chaldean catholic church, 209. us, 210. standing conference of orthodox bishops in america, 211. tikhon of moscow, 212. ordination, 213. antiochian orthodox archdiocese of north america, 214. saint cyril, 215. justinian i, 216. gregory palamas, 217. spain, 218. belarusians, 219. orthodox church of antioch, 220. hagia sophia, 221. ecumenism, 222. ecumenical patriarch, 223. hesychasm, 224. great moravia, 225. new rome, 226. albanian orthodox church, 227. autocephalous, 228. theosis, 229. transubstantiation, 230. glagolitic alphabet, 231. communion, 232. bulgarian language, 233. immaculate conception, 234. pelagianism, 235. 313, 236. the virgin mary, 237. worship, 238. macedonian orthodox church, 239. church of greece, 240. old believers, 241. mark the evangelist, 242. bolshevik revolution, 243. finnish orthodox church, 244. armenian catholic church, 245. ukrainian greek catholic church, 246. orthodox church of cyprus, 247. purgatory, 248. kievan rus, 249. 1937, 250. atheist, 251. 2000, 252. church fathers, 253. fasting, 254. edict of milan, 255. polish orthodox church, 256. ruthenian catholic church, 257. georgian orthodox and apostolic church, 258. latin empire, 259. caesaropapism, 260. decades, 261. eastern orthodox church calendar, 262. judas iscariot, 263. angels, 264. sacraments, 265. ukrainian orthodox church, 266. history of europe, 267. ukrainian autocephalous orthodox church, 268. 886, 269. jesus prayer, 270. 893, 271. eparchy of krizevci, 272. ecclesiastical latin, 273. melkite greek catholic church, 274. ecclesiology, 275. pejorative, 276. templon, 277. history of the balkans, 278. ethiopian orthodox church, 279. ecumenical councils, 280. australia, 281. catherine of alexandria, 282. psalm, 283. japanese orthodox church, 284. ukase, 285. orthodox ohrid archbishopric, 286. serb orthodox, 287. church of the genuine orthodox christians of greece, 288. czech and slovak orthodox church, 289. chinese orthodox church, 290. orthodox church of mount sinai, 291. dhimmi, 292. revised julian calendar, 293. pro-life, 294. pope benedict xiv, 295. romanian greek-catholic uniate church, 296. rus people, 297. herman of alaska, 298. 40, 299. 662, 300. fourth ecumenical council

PageRank metric reveals an overwhelming and somewhat astounding dominance of concepts tightly related to religion, which is only more surprising if one considers how closely connected seem to be most of the first 300 entries of the PageRank, with respect to the large variety of HITS' authority ranking. This undoubtedly shows how inherently different aspects of the underlying structure of Wikipedia the two metrics quantify.

4.3 Countries and Cities

This is the ranking for political geographical names.

> **HITS Authority:** United States, France, United Kingdom, Germany, Canada, England, Australia, Japan, Italy, Europe, Spain, Russia, India, Netherlands, Sweden, London, Poland, Mexico, Belgium, Soviet Union, Switzerland, Austria, China, New Zealand, Greece, Norway, Paris, Denmark, New York City, California, Portugal.

Notice that *United States* is the most authoritative country, but the most authoritative American city, *New York City*, comes after *London* and *Paris*.

> **PageRank:** United States, Russia, Rome, Canada, Europe, Japan, Constantinople, United Kingdom, France, Greece, Egypt, China, Germany, Syria, England, Middle east, Moscow, Eastern Europe, India, Soviet Union, New York City, North America, Poland, Sweden, Asia, Austria, Netherlands, London.

Pagerank results seem to transcend from recent political events to privilege a wider historical and cultural perspective in weighting geographical entities. Curiously enough, it is also remarkably less Western-centric and more "global" in its classification than HITS.

4.4 Historical Events

> **HITS Authority:** World War II, World War I, 2004 Summer Olympics, 2003 invasion of Iraq, September 11, 2001 attacks, American Revolutionary War, War of 1812, Spanish Civil War, Spanish-American War, Franco-Prussian War, 1936 Summer Olympics

> **PageRank:** World War II, September 11, 2001 attacks, East-West Schism, World War I, Sino-Japanese War 1937-1945, Chinese Civil War, Second Vatican Council, Sook Ching Massacre, Kaimingye Germ Weapon Attack, Nanking Massacre, Cold War, Russian Revolution

The trend described in the previous section seems confirmed here: Pagerank shows a more global perspective over world historical events, while HITS authority highlights a strong bias toward America-related events.

4.5 People

> **HITS Authority:** George W. Bush, Adolf Hitler, Bill Clinton, Ronald Reagan, Saddam Hussein, Winston Churchill, Richard Nixon, John F. Kennedy, Elizabeth II of the United Kingdom, Tony Blair, George Washington, Thomas Jefferson, Abraham Lincoln, Jimmy Carter, Woodrow Wilson, William Shakespeare, The Beatles, Theodore Roosevelt, Osama bin Laden, Karl Marx, Joseph Stalin, Benito Mussolini, Pope John Paul II, Aristotle, Jesus, Benjamin Franklin, Dwight D. Eisenhower, Harry S. Truman, Fidel Castro, Donald Rumsfeld.

PageRank: Jesus, Paul of Tarsus, Saint Peter, Aristotle, Tertullian, Constantine I of the Roman Empire, Church Fathers, Pope Benedict XIV, Mao Zedong, Pope John Paul II, George W. Bush, Martin Luther, Muhammad, John Calvin, George Washington, John Wesley, Karl Marx, Chiang Kai-Shek, Plato, Abraham Lincoln, Alexander the Great, Thomas Jefferson, Donald Trump, Thomas Aquinas, Bill Clinton, Morgan Stanley, Elizabeth II of the United Kingdom, Adolph Hitler, Lewis Mumford, Saddam Hussein, Ronald Reagan, Minoru Yamasaki, Christopher Columbus, William Shakespeare.

HITS's authority reveals a relatively strong bias towards recent political leaders, while Pagerank scores at higher levels several characters with an impact on religion, philosophy and society. One cannot but notice that no woman occurs in any of the lists.

4.6 Common Nouns

HITS Authority: Television, Scientific Classification, Animal, Population, Kilometre, Bird, Film, Radio, Automobile, Internet, Top-level domain, Law, Jazz, Religion, Economics, Government, Football (soccer), Species, Music, Boxing, Democracy, God, Philosophy, Science Fiction, Plant, Gold, Terrorism, Politician, Constitution, Politics, Mathematics, Language, Science, Tennis, University, Slavery

PageRank: Pope, God, Priest, Salvation, Faith, Monastery, Altar, Cathedral, Schism, Icon, Monk, Dogma, Tzar, Atheist, Religion, Biography, Manga, Christian, Saint, Law, Kilometre, Internet, Computer, Population, Economics, Language, Capital, Government, Currency, Mathematics, Film, Area, Philosophy, Politics, War Crimes, Atrocity, Revisionist, Wiki, Metre, History, Prayer, Sacrament, Baptism, Operating System, Science, Constitution, Democracy, State

Once again PageRank reveals a strong bias towards religion related issues. It would be worth analyzing what kind of cultural bias is grasped by a metric that scores at the three highest places "Television, Scientific Classification, Animal", with respect to one which places at the three top positions "Pope, God, and Priest".

5 Conclusions and further work

This first tentative work revealed some interesting facts about the cultural biases underlying the overall structure of Wikipedia. Not surprisingly it seems clear that it is a resource strongly biassed towards Western culture and history. In this sense it will be interesting to apply the very same experiment to every local Wikipedia resource online to highlight local cultural, historical or political biasses. Furthermore, although superficially similar,

Pagerank and HITS algorithm seem to grasp whole different classes of concepts that deserve further analysis.

One further step in the analysis of these first raw results will be to compare and classify them by means of additional "semantic" information. As a first stab Wordnet categories of hyponimy and hypernimy could be used to cluster the results and display quantitative properties with respect to the two chosen metrics.

We believe that, in spite of the conceptual simplicity of the experience we lead on Wikipedia data, these first results show undoubtedly how network analysis tools can be successfully applied to provide some non trivial, quantitatively grounded insights into such a particular and interesting artifact of human knowledge.

References

[1] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[2] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.