Bias In Wikipedia: Different Links, Different Stories

Raine Hoover
raine@cs.stanford.edu
CS 229 Final Report: December 16th, 2016

Abstract

We attempt to uncover the latent biases in different narratives on Wikipedia by investigating the link structure of the within-Wikipedia link graphs generated from a seed article of interest. The case study presented in this report is the Wikipedia article entitled "Arab-Israeli Conflict". We perform three experiments on this dataset: 1) hierarchical clustering on personalized PageRank vectors for each of the largest languages on Wikipedia, 2) logistic regression classification of an article as 'Hebrew' or 'Arabic' based solely on the article's links, 3) principal component analysis of the articles written in Hebrew and Arabic. We find that while clustering and PCA are inconclusive with regards to language being a primary explanation of variance between link structures, we can accurately classify the language of an article based on its concept links.

1 Introduction

There are many different historical narratives for the same underlying set of historical events. For instance, when examining the Israeli-Palestinian conflict, there are two dominant narratives: the Pro-Palestinian side tells a story in which immigrant Jews systematically displaced and oppressed native Palestinians, while the Pro-Israeli side presents a story of an oppressed people returning to their homeland. How can the same facts lead to such different narratives, and how can we relate multiple narratives to each other?

Wikipedia affords a unique opportunity to study these questions. Each article topic is resolved across all languages via the project Wikidata (Wikidata, 2016), so each language can serve as a proxy for a narrative surrounding that topic. We examine the within-Wiki link structure for different seed articles, looking for patterns across languages that correspond to geopolitical alignments around historical events.

Specifically, for the networks generated from the seed article "Arab-Israeli Conflict", we run: 1) hierarchical agglomerative clustering on each language's personalized PageRank vector starting from the seed article, 2) logistic regression on the adjacency matrices for Hebrew and Arabic, classifying whether a given article is written in Hebrew or Arabic based on which concepts it does or does not link to, 3) PCA on the adjacency matrices for Hebrew and Arabic, investigating the main axes of variance across individual article link structure.

2 Related Works

There has been some research into the differences in link structure across languages on Wikipedia. In particular, the project Omnipedia (Bao et al., 2012) highlights these differences, but leaves analysis to the user.

With regard to cultural bias in Wikipedia, much of these investigations have been descriptive and/or limited to case studies, focusing for instance on coverage of famous people in English and Polish (Callahan and Herring, 2011). Another study by Laufer et. al. focuses on several case studies, examining how different European cultures perceive one another's food practices. The authors use variations on the Jaccard similarity between the inter-Wikipedia links of the same article across languages to estimate the level of "cultural affinity" and "cultural understanding" between two European cultures (Laufer et al., 2015). While this study makes use of the within-Wikipedia link graph, it does not take into account the structure of these networks as we do in the PageRank clustering task, and looks only at pairwise comparisons between articles' link structures across languages. Furthermore, while it seeks to quantify agreement and understanding between cultures, we hope to delve into the underlying sources of cultural difference and distance.

Hecht et. al. adopts an approach most similar to the current work, looking at the network structure of the within-Wikipedia link graph across different languages to investigate a form of bias called self-focus, the tendency of a language to emphasize the home region of that language (Hecht and Gergle, 2009). While Hecht et. al. uses similar methods to study the nature of different languages' Wikipedias, the current study goes deeper in asking how languages with very close

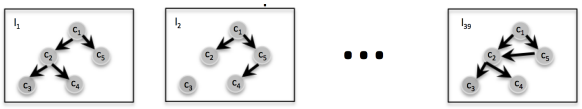

Figure 1: Example graphs– different link structures over the same set of concepts $(c_1, c_2...c_5)$ in different languages' Wikipedias $(l_1, l_2...l_{39})$.

or overlapping home regions diverge in their coverage or emphasis on particular concepts dealing with these home regions.

3 Dataset and Graph Construction

Wikidata is an organization that maintains a mapping from every Wikipedia article in every language to a resolved ID for that topic across all Wikipedias, i.e.: "The Six Day War" at en.wikipedia.org and "Guerre des Six Jours" at fr.wikipedia.org both refer to the resolved ID Q49077 (Wikidata, 2016). They provide a dump of their mappings, which we processed to get a mapping from the concatenated language code with the article title to the resolved ID for this article topic. The dump used for this milestone is slightly out of date (from November 2015), but the holes were negligible– on average, $\sim .01\%$ of links discovered in any given article were not present in the mapping generated from this stale dump.

We then began with the seed article previously mentioned, "Arab-Israeli Conflict", and queried the Wikipedia API using pywikibot (Pywikibot, 2016) to get the links for this article, which we then mapped to their resolved IDs using the mappings from Wikidata. If we encountered a link that did not have a resolved ID in Wikidata's mapping, we first resolved any redirects this link may lead to, and attempted to find a resolved ID for the article title of the redirect. We repeated this sequence in a depth first search algorithm, allowing for a depth of 2 hops, to construct the resolved inter-wiki link graph for each language (see Figure 1).

3.1 Data for PageRank Clustering

While we constructed graphs for all 75 languages that have articles on the Arab-Israeli conflict, we were only interested in clustering the largest 45 Wikipedias, a list obtained from Wikipedia's reporting of its Wikis' size (Wikipedia, 2016). Our interest in larger Wikipedias stems from the fact that a larger Wikipedia implies a more trafficked site and a more impactful narrative; in addition, a larger Wikipedia could mean more thorough coverage and a richer graph to study. Of these 45 languages, we were able to obtain a resolved ID for the seed article "Arab-Israeli Conflict" and thereby construct a link graph for 39. We report on the size of these graphs in Table 1. In total, we examine 39 languages with 207,375 resolved article IDs (concepts). Since we were aiming to model what a reader in a particular language would see were he or she to surf through Wikipedia's offerings starting from the seed article, we ran personalized PageRank from the seed article in each languages' network. Thus for our clustering task we were working with the matrix A, where A_{ij} = the PageRank for concept (resolved article ID) j in language i's concept graph.

3.2 Data for Logistic Regression and PCA

For our classification work and for PCA, we limit the languages we are investigating to just Hebrew and Arabic, and examine the adjacency matrix of the articles and concepts in the two languages. That is, we work with the matrix B, where:

$$B_{ij} = \begin{cases} 1 & \text{if article i has a link to concept j} \\ 0 & \text{if otherwise.} \end{cases}$$

An article i can either be in Hebrew or Arabic, and a concept j is resolved across both languages. Of the 54,328 total articles we examine, 28,502 are from the Arabic Wikipedia and 25,826 are from the Hebrew Wikipedia. Across both languages, there are 47,521 concepts. Of these concepts, 19,412 have articles both in Hebrew and Arabic that can be cited (as opposed to concepts that only appear as either Hebrew or Arabic articles). We run our experiments both on this "intersect" graph limited to concepts with articles in both languages and on the graph with all 47,521 concepts, and note any differences in outcomes.

For the classification task, we take a stratified split of the data between training and test sets. In the test set, there are 14,174 Arabic articles and 12,990 Hebrew articles, for a total of 27,164 articles. In the training set, there are 14,328 Arabic articles and 12,836 Hebrew articles, for a total of 27,164 articles in the training set.

For PCA, before running the algorithm we center the data: we make the mean of each feature column 0, and the standard deviation of each feature column 1.

4 Methodology

4.1 Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering is an unsupervised learning algorithm that iteratively groups samples in a bottom-up fashion, merging those samples or groups of samples that are closest to each other by some distance metric until a set number of clusters (usually one) remains. Hierarchical clustering in general is most useful for cases where the number of clusters is unknown, in contrast to k-means clustering where the number of clusters must be specified (Manning et al., 2008). There are a variety of algorithms for hierarchical clustering, each of which defines the distance between two clusters differently. In accordance with our aim of finding the differences between languages' link structures, we focus on complete-link clustering, in which the similarity between clusters is defined as the distance between their two most distant members. We experimented with a variety of distance metrics, but as we want to examine the similarity between the PageRank vectors regardless of their magnitudes, we focus on

Language	Nodes	Edges
Polish	18888	45231
Basque	1598	2550
Croatian	9366	32406
Indonesian	7344	16500
English	65450	207319
Dutch	7033	13349
Hungarian	14589	53289
Serbo-Croatian	12805	51590
Slovenian	748	1064
Catalan	11898	37439
Korean	5994	16561
Armenian	421	597
Romanian	1017	1500
Persian	12919	56591
Portuguese	16357	44086
Hebrew	25826	162188
French	19341	45868
German	22975	49185
Czech	2671	3444
Slovak	7031	24243
Norwegian	8568	20022
Danish	3879	10319
Vietnamese	5243	15297
Finnish	7677	16217
Russian	37503	133417
Spanish	27814	62825
Ukrainian	15869	48485
Turkish	11194	44499
Swedish	2616	4174
Kazakh	959	2647
Chinese	15546	41233
Esperanto	6691	12612
Malay (macrolanguage)	3050	9311
Estonian	3291	8664
Japanese	33547	84882
Bulgarian	6505	14412
Serbian	6314	21288
Arabic	28502	153589
Waray (Philippines)	798	1278

Table 1: Sizes of the within-Wikipedia link graphs for each language.

cosine similarity, defined as (Manning et al., 2008):

$$sim(a, b) = \frac{a \cdot b}{|a||b|}$$

4.2 Logistic Regression

Logistic regression is a supervised learning algorithm that uses the sigmoid function to constrain the output of linear regression– in which features are weighted by coefficients and summed to get a predicted value– to values between 0 and 1, in order to solve binary classification tasks. Logistic regression makes predictions of the following form:

$$h_\theta = \frac{1}{1 + e^{-\theta^T x}}$$

In order to fit θ in training, the algorithm seeks to maximize the likelihood of the data. Thus it seeks to maximize:

$$L(\theta) = \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1 - h_\theta(x^{(i)})^{1-y^{(i)}}$$

where m is the number of training examples, $x^{(i)}$ is the ith training example, $y^{(i)}$ is the label for the ith training example, and h_θ is as defined above. Stochastic gradient descent is used to update θ is the direction of fastest increase of the likelihood (equivalently, log-likelihood), via the update rule:

$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}$$

4.3 Principal Component Analysis

Principal component analysis, PCA, is a method of dimensionality reduction that attempts to find the axes that most explain variance across the dataset. While it is often used as a way to improve performance by collapsing features and thereby projecting the data into a lower dimensional space, PCA can also be used to uncover latent dimensions in the data (in our case, the latent dimension we hope to discover in the link structure of different articles is the articles' language) (Niculae et al., 2015). For the first principal component, PCA finds the unit vector u that maximizes the variance over all the data points:

$$\arg\max_{u:u^T u=1} u^T (\frac{1}{m} \sum_{i=1}^{m} x^{(i)} (x^{(i)})^T) u$$

which is equivalent to finding the principal eigenvector of the covariance matrix of the data, Σ. Projecting the data into a k-dimensional subspace (that is, finding the first k principal components) is equivalent to finding the top k eigenvectors of Σ.

There are different approaches to finding these eigenvectors, one of which is running singular value decomposition, or SVD, on the centered data X, which decomposes X into matrices U, S, and V^T (Berrar et al., 2003):

$$X = USV^T$$

As mentioned in section 3.2, we centered the columns of our data by scaling the mean of each feature to 0 and the standard deviation to 1. The implementation we chose for PCA therefore uses SVD internally.

	Arabic	Hebrew
Israeli Aggression		
King David Hotel Bombing	0.00071	5.85E-06
1948 Palestinian Exodus (Nakba)	0.00079	8.17E-05
Plan Dalet	0.00056	3.96E-06
Palestinian Aggression		
1929 Hebron massacre	0.00017	0.00060
Terrorism	5.39E-05	0.00089
2014 kidnapping and murder of Israeli teenagers	2.25E-06	0.00061

Table 2: Differences in PageRank for controversial (violent) events in the history of the Arab-Israeli conflict

5 Results And Analysis

5.1 PageRank Clustering

The resulting clustering using complete clustering with the cosine similarity metric can be seen in Figure 2. The left-branching nature of the clustering suggests that the subspace is too large for meaningful clusters to appear: a majority of the languages' PageRank vectors are spread throughout the subspace nearly uniformly. Furthermore, there is no clear grouping across languages– geography, language, and political alignment all fail to explain the groupings of languages at the leaves of the dendrogram. The two languages that we most expect to diverge in narrative, Hebrew and Arabic, are among the closest to each other. Indeed, the correlation coefficient between the Hebrew and Arabic PageRank vectors is .997.

Yet, inspecting the PageRank vectors for these two languages for known controversial events in the history of the conflict reveals an interesting pattern, shown in Table 2: for both Hebrew and Arabic, acts of aggression perpetrated by people who speak that language are less heavily weighted than acts of aggression perpetrated by people who do not speak that language. That is, both languages place more weight on times that their speakers were victims of violence, and less weight on times that their speakers were perpetrators of violence. The effect is more pronounced for Hebrew as opposed to Arabic, perhaps because there is only one country where a majority of citizens speak Hebrew, and therefore that language's Wikipedia more closely mirrors that country's dominant narrative. This key phenomenon is getting drowned out by the sheer number of non-zero PageRanks that Hebrew and Arabic have in common, in contrast to the rest of the languages in the subspace.

5.2 Logistic Regression

The discovery of differences in PageRank between Arabic and Hebrew motivated the dive into classification of an article's language based solely on its links. After training our logistic regression classifier, we evaluated its performance on the test set (described in section 3.2). We achieve 95.325% accuracy on the full concept matrix, and 93.377% accuracy on the "intersect" matrix. The confusion matrix for the classifier is given in Table 3. In both, the accuracy for Arabic is higher than that for Hebrew, likely because there are ∼1,500 more Arabic articles than Hebrew articles in the training set.

	'ar'	'he'
'ar'	13,715	459
'he'	811	12,179

(a) Full

	'ar'	'he'
'ar'	13,755	500
'he'	1,299	11,610

(b) Intersect

Table 3: Confusion matrices for the logistic regression classifier when using either the full list of concepts (3a) or only concepts that appear in both Arabic and Hebrew (3b).

The high performing classifier is a good indication that we can identify a particular language's narrative based solely on the concepts an article uses to explain a topic. In addition, an examination of the coefficients (θ) learned by the classifier aligns with the differences in PageRank noted between controversial (violent) article topics in Arabic and Hebrew, as well as illuminates other key differences in the framing of the concepts related to the conflict. Salient categories for the concepts that were heavily weighted towards Hebrew included:

- Arab aggression: "Palestinian political violence", "Palestinian stone-throwing"
- Loaded pro-Israel terms: "Land of Israel", "Aliyah", "State of Palestine"
- Nuclear threat: "Plutonium", "Nuclear reactor"

Salient categories for the concepts that were heavily weighted towards Arabic included:

- Israeli aggression: "1982 Lebanon War", "1948 Palestinian Exodus"
- Religion: "Hebrew calendar", "Book of Ruth", "Jewish philosophy", "Islam", "Christianity", "Judaism"

Inspecting θ also suggested that the classifier was picking up on differences between the languages that did not have to do with narrative framing, but rather idiosyncrasies in Wikipedia practices across the two languages. For instance, the Hebrew Wikipedia tends to link to years and dates much more than the Arabic Wikipedia.

5.3 Principal Component Analysis

The effects on our experiments of these idiosyncrasies across different languages' Wikipedias can be seen even more strongly in our PCA results. As seen in Figure 3, the data does cluster along the first primary component, though not clearly according to language. Investigating the ordering induced on the articles by this component suggests a distribution over topic, with the rightmost cluster being a cluster of all the years and dates. The second principal component seems to group articles into geographical places.

The variance explained by any one principal component is small, with the first 10 principal components explaining ∼11% of the variance in the data, and the first 50 principal components explaining ∼32%. This implies that the concepts are fairly independent– they cannot be easily collapsed into very few components

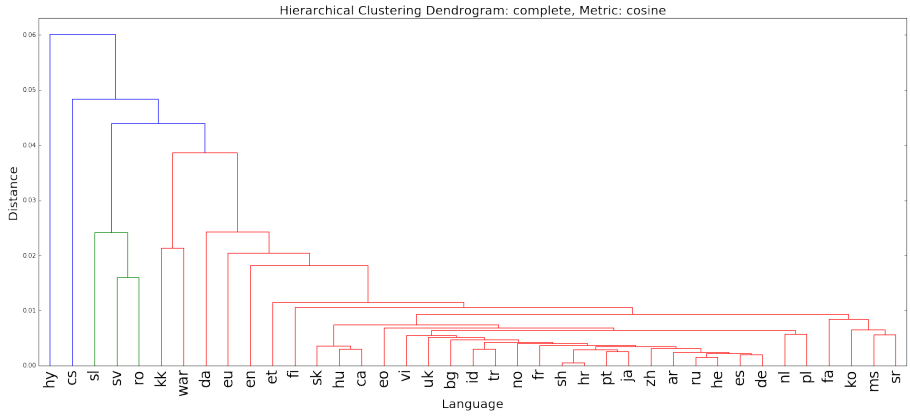

Figure 2: Output of hierarchical agglomerative clustering using the 'complete' algorithm and the cosine similarity distance metric on the personalized PageRank vectors for each languages' concept graph, staring from the seed article "Arab-Israeli Conflict". Languages are labeled using ISO 2 Letter language codes.

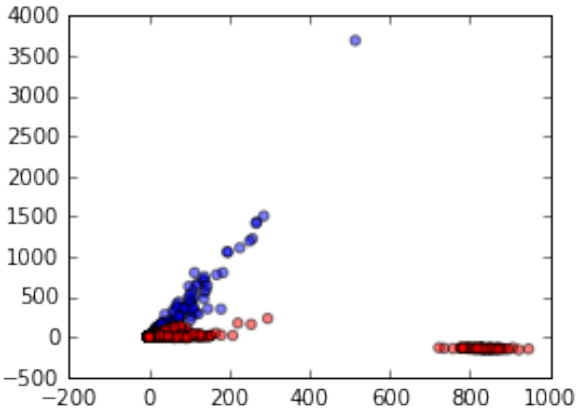

Figure 3: Plot of the data along the first two principal components. Hebrew articles are marked in red, Arabic articles in blue.

that explain a majority of the variance of the data. 92% of the variance is explained by the first 500 components, however– given that we began with a space of ∼50,000 concepts, dimensionality reduction is still effective.

It appears that the most variance in the data is not aligned with language, though it is clear that Arabic and Hebrew articles do cluster in different subspaces (as the logistic regression confirmed). We would ideally like to find a component that captures the variance in narrative, and see that Hebrew and Arabic are clustered distinctly along this component. However, PCA focuses on those components that most explain variance– the experiment picked up on characteristics that did not directly relate to the narrative and thus did not group articles according to their narratives (languages).

6 Conclusion and Future Work

Overall, the experiment yielded mixed results: clustering and PCA did not yield obvious endorsements of our hypothesis, but the high performance of the logistic regression classifier suggests that indeed we can define a particular languages narrative based on the concepts an article uses to explain a topic. Furthermore, the coefficients θ learned by the classifier endorse the theory that specific events are emphasized or deemphasized in producing a given narrative.

Both the PageRank clustering and the PCA experiments suggest an informed reduction in feature space could lead to more clearly conclusive results. The differences in PageRank between Hebrew and Arabic were drowned out by the sheer number of concepts included, many of which were irrelevant. The components that PCA picked up on were not components that one would expect to differ across language (i.e. topic and geographic location). This suggests that a useful next step for the project would be to prune the concept graphs that form the basis for these experiments, perhaps using a Jaccard similarity threshold between the article that is linking to a concept and the article that represents that concept in that language.

A more detailed error analysis of the logsitic regression classifier could also prove useful for this task: perhaps articles that the classifier came close to misclassifying (those whose predicted probabilities were close to .5) could be considered neutral concepts, and used as a way to limit the inclusion of articles irrelevant to the divergent narratives. Similarly, simply pulling out the concepts whose articles were close to equally likely to be Arabic as Hebrew could provide a source of neutral articles to be excluded from the studies.

References

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1075–1084, New York, NY, USA. ACM.

Daniel P Berrar, Werner Dubitzky, Martin Granzow, et al. 2003. *A practical approach to microarray data analysis*. Springer.

Ewa S. Callahan and Susan C. Herring. 2011. Cultural

bias in wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915.

Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on Communities and technologies*, pages 11–20. ACM.

Paul Laufer, Claudia Wagner, Fabian Flöck, and Markus Strohmaier. 2015. Mining cross-cultural relations from wikipedia: A study of 31 european food cultures. In *Proceedings of the ACM Web Science Conference*, WebSci '15, pages 3:1–3:10, New York, NY, USA. ACM.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*, pages 798–808. ACM.

Pywikibot. 2016. pywikibot 2.0rc5 : Python package index. `https://pypi.python.org/pypi/pywikibot`.

Wikidata. 2016. Wikidata:main page. `https://www.wikidata.org/wiki/Wikidata:Main_Page`.

Wikipedia. 2016. List of wikipedias. `https://en.wikipedia.org/wiki/List_of_Wikipedias#Detailed_list`, Nov.