The Role of Geographic Information in News Consumption

Gebrekirstos G. Gebremeskel
CWI, Amsterdam
gebre@cwi.nl

Arjen P. de Vries
CWI, Amsterdam
arjen@acm.org

ABSTRACT

We investigate the role of geographic proximity in news consumption. Using a month-long log of user interactions with news items of ten information portals, we study the relationship between users' geographic locations and the geographic foci of information portals and local news categories. We find that the location of news consumers correlates with the geographical information of the information portals at two levels: the portal and the local news category. At the portal level, traditional mainstream news portals have a more geographically focused readership than special interest portals, such as sports and technology. At a finer level, the mainstream news portals have local news sections that have even more geographically focused readerships.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Practice

Keywords

Geographic Information; News Consumption; News Recommendation

1. INTRODUCTION

Online news reading is increasingly becoming the norm, with traditional newspapers moving to online service provision and new news portals and aggregators emerging. One problem with online news provision and consumption is the overwhelming number of news items available to consumers. It is in the interest of news providers and news consumers to mitigate this overload. This has resulted in the emergence of news recommender systems, systems that attempt to solve the overload by proactively recommending the news items that are deemed interesting to the news reader. The success of a recommender system depends on the understanding of the factors that affect news consumption. This includes understanding both the content of the news items and the behaviors and preferences of news consumers.

These factors can be categorized into content and non-content. Content factors are modeled by key-words and named entities [1], and topics [4]. Non-content factors, include, among others, the user's current context, social media annotations and other subtle features. Social media annotations affect both user's news consumption and satisfaction [3]. Branded companies and friend annotations and recommendations increase both consumption and satisfaction [3]. The subtle features such as readability, writing style, the type of a story, visual complexity, and use of photography also influence a user's decision to read a news item [2]. It has been shown that non-content factors are as competitive as content-based factors in influencing the user's decision to read news items [2]. However, to the best of our knowledge, we have not seen a study on the effect of geographical proximity on news consumption.

This study investigates the role of geographic information in news consumption. It is a descriptive analysis work with the goal of assessing the role of geographic information in news consumption and see its potential for news recommendation. Recently, item recency has been shown to be an important factor in news recommendation [5]. Together with geographic information, this spatio-temporal features may be attractive because they are easy to implement and computationally efficient.

Using a dataset of user-news item interactions during a one-month period, we analyze and quantify the role of users' and items' geographical information in news consumption. Analysis is done at two levels: the information portal level and the local news category level. The contributions of the paper are as follows: 1) Analysis and comparison of information portals based on geographical distribution of their news readers. 2) Investigation of the local and non-local news categories of mainstream portals with respect to the geographical distribution of their readership and 3) Describing and quantifying the role of the relationship between geographic information of mainstream information portals (and their local and non-local categories) and user's geographic location on news consumption.

2. DATA

We use 53 million user-item interactions with items of 10 information portals collected by Plista[1], over a period

[1] http://orp.plista.com/documentation

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741716.

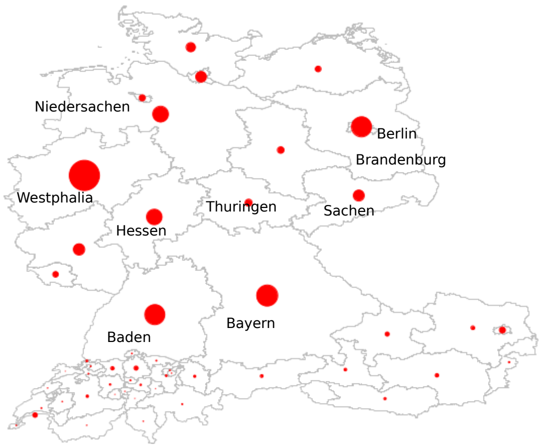


Figure 1: Bubble map of all users across states. Most users come from German states and Westphalia produces the largest number of users.

of one month. Plista provides the Open Recommendation Platform (ORP), a framework that brings together information portals and news recommendation providers (referred to as participants). When a user starts reading a news item, a recommendation request is sent to one of the participants while the other participants receive the impression information. Every participant has access to all user-news item interaction information. The logs have been annotated by Plista with URLs of news items and state-level postcodes of news consumers. From the URLs, we can detect the local news items (as opposed to the non-local news items). The state level postcodes represent the user's geographical location and the local and non-local news categories represent geography of the news items.

2.1 The Information Portals

Table 1 presents the information portals, their URLs and types. Figure 1 presents the distribution of the total number of users in our analysis by states. Most users come from Germany, and the state of Westphalia produces the highest number of news readers, consistent with the fact that it is the state with largest population. Figure 2 presents the distribution of users by portal. The automotive forum (Motor-talk), the two mainstream news portals (Ksta and Tagesspiegel) and the sport news portal (sport1) have larger readerships. Two of the ten portals (Tagesspiegel and Ksta) are traditional mainstream news portals providing opinion, politics and current events, and they can be national or regional. The other portals are special interest portals focused on information technology (4), sports news (1), automotive (1), business (1) and home and gardening (1).

2.2 Users and Items

Using cookie identifiers for user identification has a shortcoming in that, if a user does not maintain persistent account, s/he will be counted more than once. Items are identified by unique numerical identifiers. Both items and users have many attributes. For our analysis, we focus on the state-level postcodes of users and on the URLs of news items.

Table 1: The information portals. The short names are the names by which we refer to the portals in plots

Short name	Type	URL
Cfo	Business	cfoworld.de
Cio	IT News	cio.de
Woche	IT News	computerwoche.de
Gulli	IT& Games	gulli.com
Ksta	News	ksta.de
M-talk	Automotive	motor-talk.de
Channel	IT	tecchannel.de
Sport1	Sports	sport1.de
Tage	News	tagesspiegel.de
WH	Garden	wohnen-und-garten.de

2.2.1 *User Location Information: States*

Our analysis is focused on the 52 states of Germany, Switzerland and Austria for two reasons. The first reason is Plista provided us with the mapping to the real postcodes of only the three countries' proxy postcodes that are originally used to represent the different states. The real postcodes help us anchor and contextualize our findings to actual geographical locations. The second reason is that the states of the three countries are geographically close to each other, German-speaking (all the information portals are in German language) and thus of primary interest for our study.

2.2.2 *Item Location Information: (Non-)local News*

The two mainstream news portals organize their content in different sections of which city columns have our special interest, as news items deemed geographically relevant to the particular cities are placed under them. Tagesspiegel has Berlin column (http://www.tagesspiegel.de/berlin/) and Ksta has Cologne column (http://www.ksta.de/koeln/). We take advantage of the manual placement of news items (by the news editors) into the respective local news sections as a manual geotagging process. We consider all the news items that fall under a city column as local news and all the rest as non-local. We identify four different subsets to study: Berlin (**T+B**), Tagesspiegel-minus-Berlin (**T-B**), Cologne(**K+C**) and Ksta-minus-Cologne (**K-C**). For comparison, we also include Tagesspiegel's sport section (**T+S**).

3. ANALYSIS AND DISCUSSION

We analyze the information portals, with a view to finding similarities and patterns in the geographic distribution of their readership. Then we analyze the mainstream news portals and their local news categories also for similarities and differences in geographic distribution of their readerships. In both cases, we first aggregate the readers of an information portal or local categories by the 52 selected state-level postcodes. From the aggregated counts, we compute geographical likelihood distributions (across the states) of the readerships of the information portal or the local categories. Then we employ the Jensen-Shannon Distance (JSD) metric to quantify the difference between the geographic likelihood distributions (Equation 1). The uppercase letters X and Y represent vectors of likelihood distributions and KL stands for KullbackâĂŞLeibler divergence (Equation 2). As JSD is a distance metric, the smaller the distance score between two likelihood distributions, the more similar the they are. Finally, we examine and analyze how well we can correctly

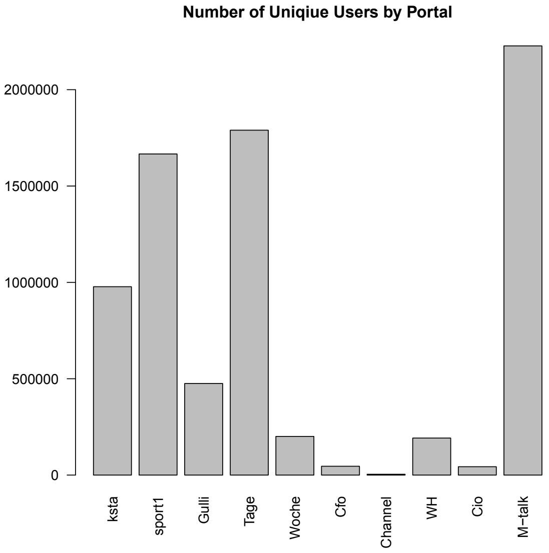


Figure 2: User frequency distribution by information portals

predict the likelihood of a user's state given the portal or the local news category the user visits.

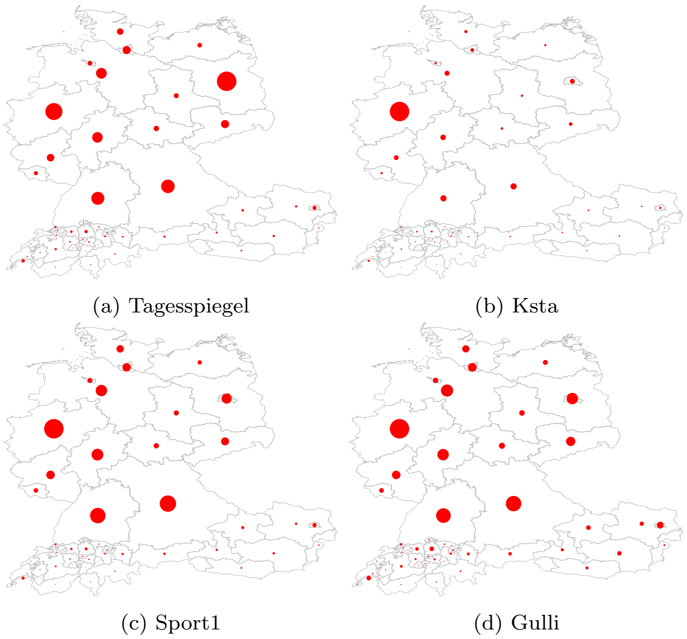
$$JSD(X,Y) = \sqrt{\frac{1}{2}KL(X, \frac{(X+Y)}{2}) + \frac{1}{2}KL(Y, \frac{(X+Y)}{2})} \quad (1)$$

$$KL(X,Y) = \sum_i x_i \ln \frac{x_i}{y_i} \quad (2)$$

3.1 Mainstream vs. Special Interest Portals

We characterize information portals by geographic distribution of their readerships modeled using conditional likelihood $P(user\ state|portal)$. Using JSD between the conditional likelihood distributions, we can determine how geographically similar their readerships are. The results are presented in Table 2. Firstly, the highest JSD observed between any two portals is 0.368, that is between Ksta and Tagesspiegel. Secondly, we observe that the first and the second highest JSD scores of every special interest portal are from Ksta and from Tagesspiegel respectively (see the colored columns and rows in Table 2) .

The first observation tells us that the mainstream news portals differ the most in geographical readership. The second observation indicates that the two mainstream news portals have geographical user distributions that are very different from those of the special interest portals. Together, these observations indicate that, even in an online world, mainstream news portals are perceived as representing a certain geographical region and their readerships are mainly from those regions, while special interest portals are not bound to a geographical region of the type mainstream news portals are. The JSD scores between each of the special interest portals are small compared to the JSD scores between special interest portals and mainstream news portals. The distance scores between every special interest portal and mainstream news portals vary from 0.187 to 0.330, whereas the distance scores between each of the special interest portals vary from 0.033 to 0.140.



(a) Tagesspiegel

(b) Ksta

(c) Sport1

(d) Gulli

Figure 3: Bubble maps of the state-level distribution of users for the mainstream news portals (Tagesspiegel and Ksta) and two specialized portals (Sport1 and Gulli). Ksta has a very localized readership. Tagesspiegel and the special interest portals show a more distributed readerships.

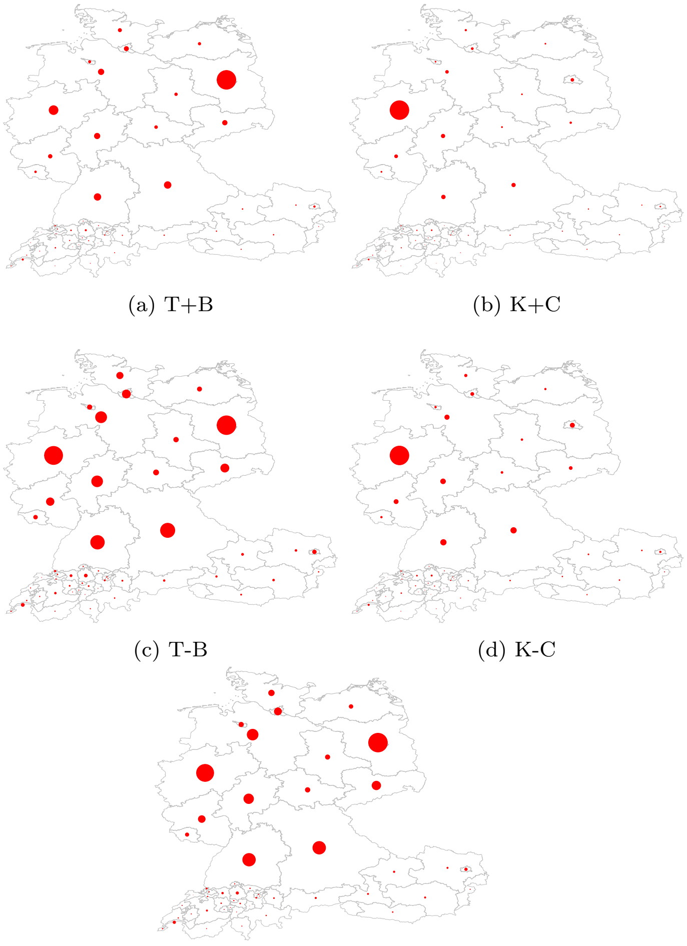
Figure 3 presents bubble maps of the user frequency counts of each state for the two mainstream news portals, and for two examples of special interest portals, for comparison. The bubble maps for the mainstream news portals have geographical foci. Ksta's readership is mainly from its homestate (Westphalia) and Tagesspiegel's readership is more distributed than Ksta's. The bubble maps for the two special interest portals (Sport1 and Gulli), however, are more evenly distributed. These observations are indications that there is an association of mainstream news portal with some geographical focus while the appeal of interest portals seems not to be limited to the same geographical constraint. The mainstream news portals are interesting for the following additional reasons too. First, they are two of the three portals that receive significant clicks on recommended articles [5]. Second, they are the portals that offer opportunity for extracting geographic local and non-local news categories, which we discuss in Subsection 3.2.

3.2 Local vs. Non-local News Categories

For each local news category, users are aggregated by state-level postcodes. Then we compute $P(user\ state|locale)$ which is a geographical likelihood distribution (across the states of the three countries) of the local news readership. Using the geographical likelihood distributions, we compute JSD scores between the four news categories. The results are presented in Table 3. The highest distance observed (0.561) is between Berlin category of local news (**T+B**) and Cologne category of local news (**K+C**), an indication that the geographical distributions of their readerships are the the most different. The next highest distance observed (0.485) is between Ksta and **T+B**.

Table 2: Adjacency matrix of information portals based on the Jensen-Shannon distance between the geographic distribution of their readerships. The highlights show the distances between special interest portals and mainstream news portals.

	WH	M-talk	Tage	Woche	Cio	Cfo	Channel	Ksta	Sport1
Gulli	0.067	0.057	0.187	0.066	0.101	0.129	0.043	0.322	0.102
Sport1	0.099	0.080	0.192	0.091	0.105	0.131	0.119	0.305	
Ksta	0.330	0.314	0.368	0.323	0.321	0.332	0.331		
Channel	0.067	0.062	0.209	0.055	0.087	0.111			
Cfo	0.140	0.127	0.229	0.082	0.053				
Cio	0.110	0.093	0.215	0.044					
Woche	0.076	0.060	0.198						
Tage	0.221	0.210							
M-talk	0.033								



(a) T+B

(b) K+C

(c) T-B

(d) K-C

(e) T+S

Figure 4: Bubble maps for the state-level distribution of the readership of the news categories of the two mainstream news portals. Note that Tagesspiegel's Berlin local news section has very different geographical readerships from Tagesspiegel-minus-Berlin and Tagesspiegel' sport section. Note also that Cologne and Ksta-minus-Cologne have almost the same geographical readership

Table 3: Adjacency matrix for the local and non-local news categories based on Jensen-Shannon distances between the geographic distributions of their readership. Note the largest distances between K+C and T+B, and between T+B and Ksta. Note also the large difference between the distance scores of T-B and T+B (0.230), and between K-C and K+C (0.133.

	Tage	Ksta	T+B	K+C	K-C	T-B
T+S	0.038	0.360	0.207	0.465	0.358	0.046
T-B	0.031	0.354	0.230	0.465	0.351	
K-C	0.366	0.003	0.483	0.133		
K+C	0.474	0.130	0.561			
T+B	0.200	0.485				
Ksta	0.368					

It is interesting to examine the differences between the different categories of news items published in the same portal. This means the distances between Tage, **T+B**, **T-B**, **T+S** on one hand, and Ksta, **K+C** and **K-C** on the other. The distance between **T+B** and Tagesspiegel is 0.200, whereas the distance between **T+S** and Tagesspiegel is 0.038. Clearly, this shows how geographically different the readership of the Berlin category of local news is from the readership of the full portal or its sports' section. The distance between T+B and T-B is 0.230, indicating further that the Berlin category of local news has a geographically more distinct readership. It is also important to compare the difference with the distance between Cologne and Ksta-minus-Cologne which is 0.133. The almost double distance between the local and non-local sections of Tagesspiegel is an indication that the Berlin category of local news and Tagesspiegel-minus-Berlin have a large difference in geographic readership distributions. We explain this by the fact that Tagesspiegel has a wider readership that covers a larger geographical area. We interpret the smaller distance between K+C and K-C as evidence that Ksta reaches a narrower geographical readership anyway, that is that Ksta has a more regional character.

Our explanation of viewing Tagesspiegel as a national newspaper as opposed to Ksta as a regional one is supported by the bubble maps of figures 4 and 3. In Figure 4, we clearly see that the readership of Berlin category of local news and Cologne category of local news are more geographically localized than those of Tagesspiegel-minus-Berlin and Ksta-minus-Cologne, and that the readership of the Cologne category of local news is more localized than that of the Berlin category of local news.

3.3 Likelihood of a User's State Given a Portal or a News Category

Another way to look at the relationship between a user's geographic location and a geographic information of portals and categories is to compute the likelihood of correctly predicting the user's state given the information portal (or category) and a cutoff value of user's visit frequency. Specifically, we compute the likelihood $P(\mathit{user\ state}|\mathit{portal}, \mathit{cutoff})$ and $P(\mathit{user\ state}|\mathit{locale}, \mathit{cutoff})$. For the special interest portals, since their readerships are geographically distributed, the likelihood of predicting a user's state correctly is very low (less than 0.2). Therefore the likelihood of predicting a user's state from their visits of portals and news categories is interesting only for the mainstream news portals and their local and non-local news categories.

For the mainstream portals, the likelihood that a visiting user is from the state of their geographical focus is very high (as compared to the likelihood that the user is from any of the other states). Therefore, we focus on the likelihood of the respective geographical focus for each of the mainstream news portals and their local and non-local categories. The results are presented in the plots of Figure 5. We observe that the likelihood that a user reading a Cologne category of local news is from the state of Westphalia is as high as 0.8, as compared to a user reading Ksta which gives the likelihood of 0.40. In the case of Berlin category of local news, the likelihood that a user is from the state of Berlin is 0.48 and in the case of Tagesspiegel, the likelihood that a user is from the state of Berlin is 0.22.

As we increase the cutoff of the frequency of visits of the user, we observe that the likelihood of predicting a user's state increases. The gap between the plots of the Berlin category of local news, and the Tagesspiegel-minus-Berlin category is a measure of the strength of the geographical information in the local news consumption. On the categories of Ksta, however, the gap between the Cologne category and Ksta-minus-Cologne is small, indicating a more or less the same readership for the local news and the portal itself. It is also worth noting that the readership of the Cologne category is small compared to the readership of the Ksta-minus-Cologne, and has no impact on the combined plot (Ksta). Our explanation for the difference in likelihoods of predicting the respective states for Cologne category of local news and Berlin category of local news is that, by virtue of the state of Berlin being the capital, it attracts users from all over the country, more than Cologne does.

3.4 Discussion

We observe that geographical information plays an important role in user's consumption of news items of the mainstream news portals, and that it manifests itself at two levels: the portal level and the local news categories level, as can be observed from tables 2 and 3. The bubble maps of figures 3 and 4 visually confirm these observations. Geographical information at the portal level manifests itself in the sense that users associate a strong or loose geographical location to the portal itself. This finding may be useful in news aggregators (such as Google News and Yahoo! News) to identify news publishers that are geographically relevant to certain users.

The second level where geographical information manifests itself is at the local and non-local categories. Such fine-grained geographical information is useful, for example, for tailoring recommendations for the local and non-local news visitors. Together with associated geographic focus of the portal, the local and non-local categories may be used for improving news recommendation. We imagine that such geographic information can be useful in big countries where there are competing national and regional news portals.

4. CONCLUSION

We have investigated a dataset of one month of user interactions with news items of different information portals. We measured the distance based on geographical distribution of readerships between different news portals and found out that mainstream news portals and special interest portals show differences in the role geographic information plays in influencing users. While the special interest portals seem to be less geographically localized, the mainstream news portals, on the other hand, exhibit geographical foci. The mainstream news portals were further analyzed by focusing on their local news categories which also showed a more localized geographical readerships. We showed the likelihood that a user is from the home-state (the geographical focus) of the mainstream news portal can be predicted reasonably well, specially when higher cutoff values of the user's visit frequency are considered. The relationship between the geographic location of news users, and the geographic foci of mainstream news portals and their local news categories can be exploited for improving news recommendation, which we plan as our future work.

Acknowledgment

This study is financed by the COMMIT program as part of the Infiniti project. The authors would like to thank Desmond Elliott and Laura Hollink for their comments and feedback on the paper, and Alan Said for providing initial guidance on working with the data.

5. REFERENCES

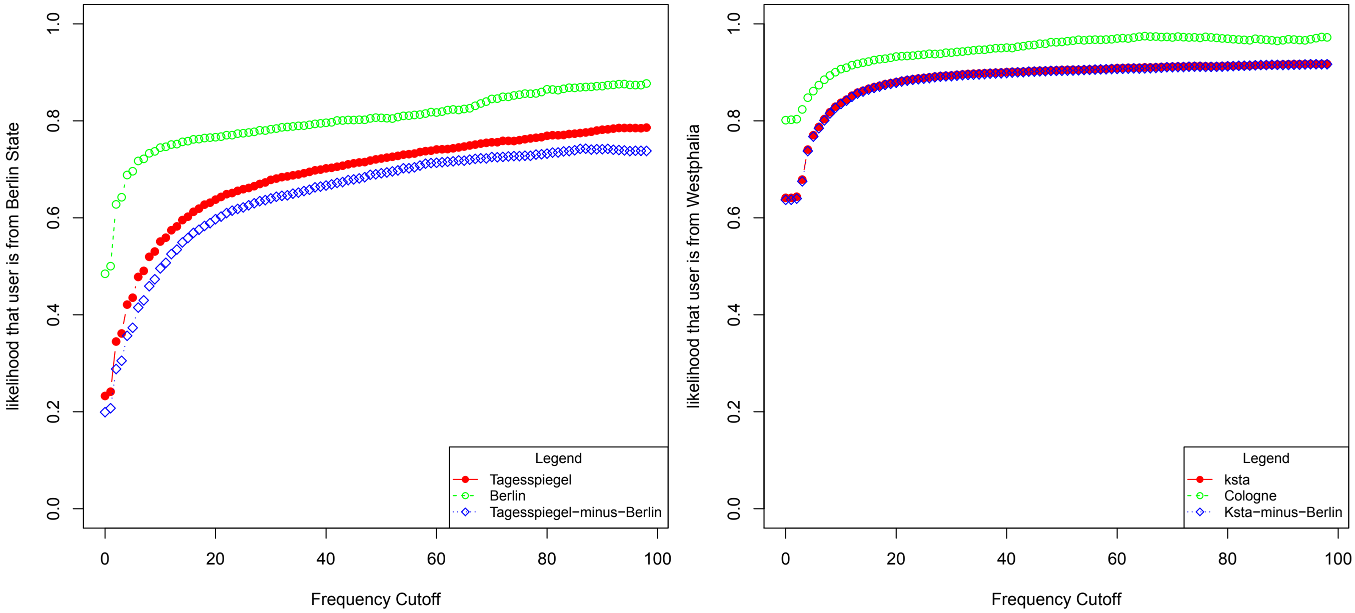
[1] Gabrilovich, E., Dumais, S., Horvitz, E.: Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In: Proceedings of the 13th international conference on World Wide Web. pp. 482–490. ACM (2004)

[2] Jancsary, J., Neubarth, F., Trost, H.: Towards context-aware personalization and a broad perspective on the semantics of news articles. In: Proceedings of the fourth ACM conference on Recommender systems. pp. 289–292. ACM (2010)

[3] Kulkarni, C., Chi, E.: All the news that's fit to read: a study of social annotations for news reading. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2407–2416. ACM (2013)

[4] Li, L., Wang, D., Li, T., Knox, D., Padmanabhan, B.: Scene: a scalable two-stage personalized news recommendation system. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 125–134. ACM (2011)

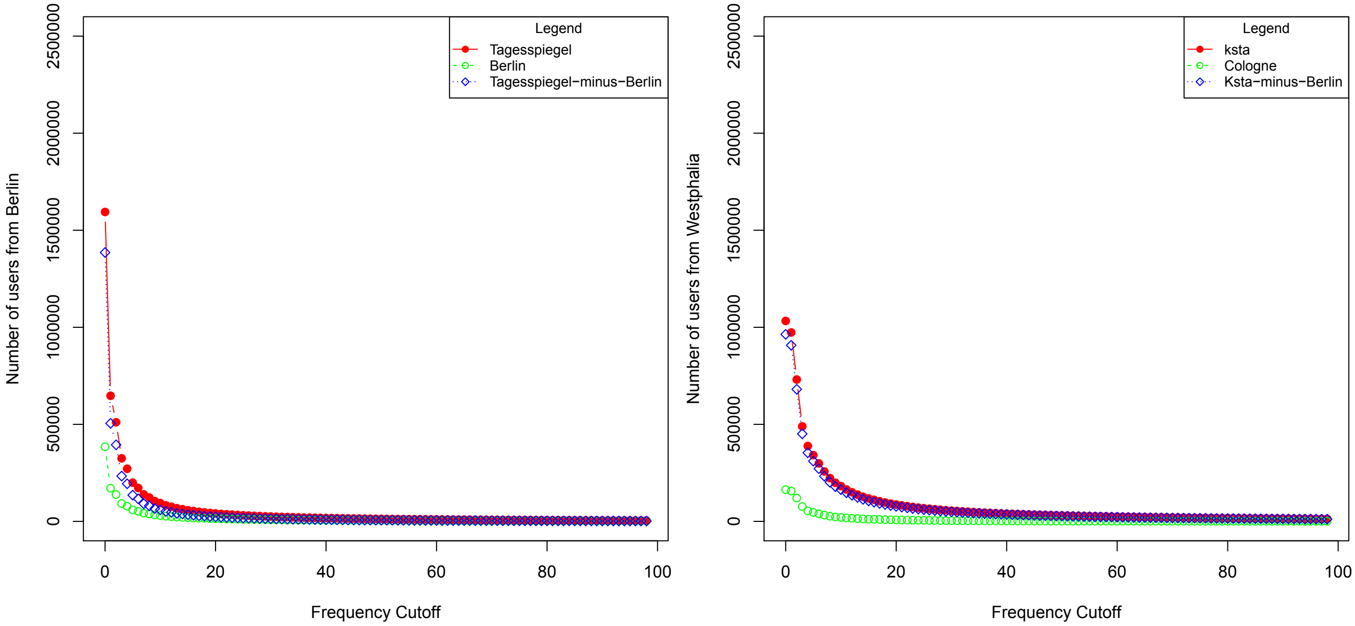
[5] Said, A., Lin, J., Bellogín, A., de Vries, A.: A month in the life of a production news recommender system. In: Proceedings of the 2013 workshop on Living labs for information retrieval evaluation. pp. 7–10. ACM (2013)



(a) The likelihoods of a user being from the state of Berlin for local and non-local news categories of Tagesspiegel

(b) The likelihoods of a user being from the state of Westphalia for the the local and non-local news categories of Ksta

Figure 5: Each figure presents the $P(Berlin\ state|locale, cutoff)$ and $P(Westphalia|locale, cutoff)$ for the news categories of Tagesspiegel (5a) and of Ksta (5b) respectively. We see a wider gap between the plots of Berlin and Tagesspiegel than between Cologne and Ksta, an indication of difference in geographical readerships of Berlin and Tagesspiegel from Cologne and Ksta. We also see that the plots of Ksta-minus-Cologne and Ksta overlap because the the number of user-item interactions for Cologne is very small compared to Ksta-minus-Cologne.



(a) The number of users for the news categories of Tagesspiegel

(b) The number of users for the news categories of Ksta

Figure 6: Each figure presents the number of users remaining versus cutoff values for the news categories of Tagesspiegel (6a) and of Ksta (6b).