Large Scale Retrieval of Social Network Pages by Interests of Their Followers

Elena Mikhalkova[1][0000−0003−0781−8633], Yuri Karyakin[2][0000−0003−2346−402X], and Igor Glukhikh[3][0000−0002−0683−6138]

[1] Tyumen State University, Tyumen, Russia
e.v.mikhalkova@utmn.ru
[2] Tyumen State University, Tyumen, Russia
y.e.karyakin@utmn.ru
[3] Tyumen State University, Tyumen, Russia
i.n.glukhikh@utmn.ru

Abstract. Social networks provide an opportunity to form communities of people that share their interests on a regular basis (circles of fans of different music, books, kinds of sports, etc.). Every community manifests these interests creating lots of linguistic data to attract new followers to certain pages and support existing clusters of users. In the present article, we suggest a model of retrieving such pages that attract users with similar interests, from a large collection of pages. We test our model on three types of pages manually retrieved from the social network Vkontakte and classified as interesting for a. football fans, b. vegetarians, c. historical reenactors. We use such machine learning classifiers as Naive Bayes, SVM, Logistic Regression, Decision Trees to compare their performance with the performance of our system. It appears that the mentioned classifiers can hardly retrieve (i.e. single out) pages with a particular interest that form a small collection of 30 samples from a collection as large as 4,090 samples. In particular, our system exceeds their best result (F1-score=0.65) and achieves F1-score of 0.72.

Keywords: Interest discovery · Social group · Major interest · Social network · Supervised machine learning.

1 Introduction and Related Work

Classifying a page as interesting or not for a user who is scrolling through a social network is not a challenge. The main issue is rather the overload of pages they have to look through before they find what they want. Hence, advancement of recommender systems that help users find communities of interest is an ongoing process characterized by a variety of approaches. The focus of these approaches is usually the user. As [13] puts it, user-modelling that generally deals with behavior and actions of a user in a computer system includes inferring interests from them (interest discovery [4]). From this perspective, one user

[4] The field is called so by [18, 39] and some other.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

can exhibit a variety of interests, and the task of modelling is to infer them. In this paradigm, the main marker of interests is linguistic data (user-generated content). Hence, interests are mined as tags [11, 18, 32, 36], keywords [4, 35, 37], named entities [3, 28, 33], user classified interests from profiles [17, 24], topics [2, 19, 20, 39] in microblogs [3, 29, 38], most commonly derived with the help of LDA and LSA algorithm [7, 34] [5]. Other approaches, e.g. the social network analysis, employ such non-linguistic information as friends, followers [12], contacts [31], clicks [1, 4], likes [8] and reposts, retweets, social recommendations [9, 10, 16]. Some projects unite users into clusters that can be represented with a graph-model [23, 40]. In all approaches, the main target is to facilitate the search functions of social networks by a more effective recommendation.

As for the algorithms of interest classification, their choice depends on the model. Where machine classification is possible, according to [25], traditionally the following classifiers are used: Decision Trees, Nearest Neighbors, Naive Bayes, linear algorithms separating hyperplanes (variations of commonly known Support Vector Machines, or SVM). [6] use Nearest Neighbors and Naive Bayes to suggest NLP-based recommendation of "news of interest". However, none of the works we know focus on community pages that atrract users with similar interests. As we demostrate below, such pages provide valuable information on existing user clusters and user interests.

In the present research, we would like to shift the focus from modelling a single user's list of interests to modelling a social network community that a user might like, and we will do it based on a linguistic model. We assume (and discuss further) that one main interest is what attracts a user to a page if they start to follow it [6].

Our solution presumes we already know a page that a user likes, or we have a set of pages that a user's friends like - we will call such pages *model*. A recommender system can find more pages that are similar to the *model* ones with the help of text similarity algorithms [7]. We can also view this task as a text classification problem usually solved with such algorithms as Decision Trees, *k*-Nearest Neighbors, Naive Bayes, etc. Additionally, classification presupposes that pages followed by users with a common interest belong to a certain *class*, especially from the sociological and linguistic point of view.

2 Interest Classification from the Sociological Perspective

Although interests are personal, in communities they have to be shared (sociologists call this phenomenon contagion [30]). In social networks, interest sharing produces linguistic content that makes online communities a valuable object of research.

[5] [27] evaluate importance of these types of linguistic content in user-modelling.
[6] Unless they already know the page owner and follow them to confirm the previously established contact.
[7] A good account of such algorithms is given by [15].

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

Although there is no universal definition of social groups, many authors among whom are [5, 14, 21], etc. agree that a social group is a collection of individuals interacting in a certain way on the basis of shared expectations of each member of the group in relation to others. A social group can be viewed as an abstract whole that has certain features distinguishing it from others. For example, football fans as a social group are known around the world for their typical behavior: attending football matches, collecting sports memorablia, and quite often for violation of public conduct. Accordingly, adherence of an individual to the social group shows in speech. An individual who claims to belong to a social group calls himself or herself by a special name (a football fan of some team, a hoolie), mentions attributes of the group (a team's name and players, leagues, places, sports memorabilia), performs activities typical of all members of the group and reports about it (attending matches, play-offs). When in social networks representatives of a social group interact, linguistic data serve as a means of identification and role assignment. Hence, network pages of social groups *can* be viewed as representatives of a class. And we can use such linguistic data as keywords, topics, named entities, terminology for automatic differentiation of these groups.

At the same time, what hinders classification is that groups can have points of intersection (for example, both football and hockey *matches* happen at *stadiums*, *teams* participate in *leagues*, etc.). Even names of teams and players can be the same. In such cases, fans often invent nicknames (using flag colors or mascots) to differentiate between them. Hence, linguistic content marks difference between unrelated social groups and simultaneously shows relation between allied groups.

Previously, we stated that there is one main interest that attracts users to a page. We will call it the Major Interest (MaI). The MaI is bound to the social group that joins for interaction on a social network page. If the people interacting do not belong to the same social group, they express different interests, and the MaI becomes unclear.

To study the phenomenon of MaI, we conducted a survey of the Russian social network Vkontakte (vk.com). We had to work with the Russian language as we were able to only find enough Russian-speaking experts. Vkontakte was created by Pavel Durov, who currently develops Telegram, in 2006. The network was chosen as one of the largest sources of linguistic content in Russian. In the experiment described in [22], we asked ten experts (certified and currently employed as linguists, sociologists, marketing specialists) to give their opinion on what social group manifests itself in a dialogue taken from a social network page. We instructed experts to define if authors in the sample dialogue belong to the same social group and, if yes, explain why they think so. The experts were not prompted by multiple choice answers. Three dialogues were marked correctly and unanimously as belonging to football fans, historical reenactors, and vegetarians. Two dialogues (fans of rock music and "bros") got a 50% agreement. And the control sample where people did not express adherence to one social group [8]

[8] The sample was taken from a page where people discussed a concert of Madonna that they attended or read about. Some of them expressed discontent with her religious

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

got a 90% agreement that there is no social group and that these people do not share any interests.

After the experiment we conducted automatic classification of social network pages by the three MaIs (football, rock music, vegetarianism) accross networks and languages. For each MaI in the three sets (English Twitter, Russian Twitter, Russian Vkontakte), we prepared 30 text samples downloaded from social network pages. We used several classifiers (SVM, Neural Networks, Naive Bayes, Logistic Regression, Decision Trees, and k-Nearest Neighbors) to predict the three MaIs in each set. Logistic Regression proved to be the best performing algorithm when operating on vector representations of 1,000 most frequent words (0 denoting presence and 1 - absence of a word in a text). Table 1 illustrates the result of classification; the score given is the average F1-score of five tests performed with Monte-Carlo cross-validation.

Table 1. Interclass classification of pages with supervised machine learning classifiers: F1-score. F - football, R - rock music, V - vegetarianism, T - Twitter, Vk - Vkontakte, En - English, Ru - Russian.

	Vk Ru			T Ru			T En		
	F	R	V	F	R	V	F	R	V
Logistic Regression	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.988	0.988

Generally, in this experiment we faced the efficiency of Bernoulli model of feature representation, i.e. word frequencies are not as important as their absence or presence. We also found out that human expertise is not a guarantee that a MaI will be difficult for classification. For example, Rock music and Vegetarianism were classified similarly well.

We tend to think that MaIs are more like umbrella terms to a variety of topics discussed by communities (for example, the MaI "football" encompasses matches, players, stadiums, events, ticket sales, memorablia). On the one hand, MaIs can be generalized into *types* of social groups: football fans are a type of sports fans, rock music fans are a type of music fans. Within the type, the variety of topics is quite similar (as in the case of hockey and football fans). On the other hand, MaIs can break into specific representatives, for example, rock music fans can be Metallica fans, Slipknot fans, etc.; football fans can be fans of Manchester United, Spartak, etc. The type determines the stable part of the user-generated content that relates some social groups, and representatives of a MaI are in charge of the entropy content that differentiates them from other representatives. [9]

and political views, some *vice versa* expressed admiration. [14] calls such accidental interactions "quasi-groups".

[9] Therefore, it is important to understand what kind of content a user would like to get if they are looking for pages of interest. E.g. if a football fan is looking for other

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

3 Retrieving texts with a certain MaI from a large collection

In the present research, we will describe an algorithm that is quite efficient when searching for pages with the same MaI in a collection much larger than the number of pages to be retrieved. We designed it on the grounds of interviews with the experts evaluating the texts in the experiment described above.

Every text T_i in the test set is weighed on the basis of one or two model texts united into one T_m in the training set to state its similarity to the model in every given class C_j (each class coresponds to one MaI). The weights are evaluated by the Relevance Function. The result is a list of texts that are considered to represent the same MaI. The classes are three MaIs from the experiment: football, vegetarianism, and historical reenactment.

A Model Text T_m is a text, chosen as a standard representative of a class. Ideally, it contains as many characteristic features of the class as possible [10]. The Relevance Function extracts these features for every class. Then, in every class, the Distribution function weighs all the texts in the test set and rates them choosing the top ranked as representatives of the class. Thus, every text can occur in more than one class.

3.1 Data selection

We conducted our retrieval experiment on a corpus of texts downloaded from Vkontakte. For the present analysis, we automatically searched through 20,000 VKontakte open access pages using Vkontakte API. 4,460 pages turned out to contain user-generated content of size from 1 to 100,523 words. We asked a panel of three experts (certified linguists and sociologists) to manually search through them to find texts of football fans, historical reenactors, and vegetarians. In the final set of texts, the three MaIs were represented by a different number of items. Next, we asked experts to find more pages (using recommended links, user reposts and Vkontakte search) to create a set of 30 texts in each class. We also removed all texts belonging to the three MaIs and texts with the lowest number of words from the initial corpus. All in all, our corpus contains 4,000 unclassified items ("Miscellaneous") and 30 texts belonging to each of the three MaIs (90 texts, in total). We consider the ratio between the class "Miscellaneous" and each

fans, do they need fans of a particular team? Which is usually the case of football fans. However, with the music or anime, they might be looking for more diverse communities - fans of different music bands, cartoons.

[10] Model Texts are characterised by intense communication of multiple representatives of a social group and are often quite large. But if the text is too large, it becomes noisy. Empirically we found out that selection of approximately 1,000 features is most effective. However, this observation requires more research. Also, we observed that some Model Texts provide better results than other; often using two texts instead of one is more effective. However, unlike common supervised learning algorithms, with the increase in the number of Model Texts (even up to 10-15) our algorithm becomes less efficient.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

of the other classes to be large-scale because the joint probability to retrieve a succession of 30 items of one class from 4,030 is very low: $\frac{30}{4030} \times \frac{29}{4029} ... \times \frac{1}{4001} = 6.82936273447e-78$.

Every text in the corpus of 4,090 was preprocessed to extract the following four parameter features:

1. *Key-words.* Key-words are selected from the normalized list of words of T_m based on differences in their frequency. In a list of words, ranked by their frequency, a key-word is a word with a frequency that differs by more than one from the word with the next lower rank (e.g. 4, 7, 11 is a good list of frequencies with large enough steps; 1, 2, 3 is not). This method excludes all n legomena (hapax, dis, tris, etc.) to single out the most characteristic set of keywords. The normalized list of keywords has stop-words excluded. For short texts the result is a list of 1-2 words, and up to 20-30 for long texts.
2. *Stems.* Stems are selected from the vocabulary after stemming words with the Porter stemmer. Interestingly, when we preprocessed the vocabulary with a morphological analyser, it lowered down the performance. Therefore, no preprocessing except stemming was employed. In the resulting list of stemmed words, if each stem is found more than three times, it is added to the list of stems. This procedure is based on the expert opinion that social groups not only use some words frequently, but develop a whole vocabulary with derivatives of these words: vegetables - vegan, vegetarian, vegetarianism, lacto-vegetarian, ovo-vegetarian, etc.
3. *Uniques.* Lists of stemmed words, that were collected in the stemming procedure (without frequencies), are compared to each other in all pairs of classes, and stems that are found only within one class are added to the list of uniques. These words are a kind of terminological dictionary that describes a group's uniqueness. In the interviews, the experts also stated that groups use unique words that are understandable only by the representatives of this group or have a special value within this group. But tests showed that these lists are formed not only from some inner vocabulary, but also from common-knowledge words describing group activities.
4. *Named entities.* Named entities are a natural part of a social group vocabulary, as the group shares its impression of people, places, etc. Also, names of a group's leaders unite it. To extract named entities from social network posts and comments, we wrote a simple heuristic NER-parser. We take only named entities with frequency more than three.

3.2 Relevance Function

The Relevance Function creates a list of features for each class. The number of types of features can vary in optimization. For the further analysis frequencies are not needed. In the tested version, we cut down Model Texts so that they would produce about 1,000 features in sum. Empirically, this method showed to be the most effective.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

The four lengths of feature arrays form a vector (v_1, v_2, v_3, v_4) in the 4-dimensional space, which serves as the basis for a right rectangular prism (a hyperrectangle, or a box). The volume of the box P_m (Model Box) is a model volume and can neither be superseded or be equal to 0. To avoid it, Laplace smoothing $\alpha = 1$ is applied to every vector:

$$\Theta_i = v_i + \alpha \tag{1}$$

Once the classifier parameters are found, the system proceeds to the analysis of the test set. Every text T_i in a test set is analyzed in the same way as the Model Text except *uniques*. Instead of them, a list of stems is used. Within each class, the algorithm searches for every element of the train text arrays among the elements of T_i and adds smoothing:

$$f(x_k, T_i) = \{1(true), \text{ if } x_k \in T_i, 0(false), \text{ if } x_k \notin T_i\} + \alpha \tag{2}$$

The result of evaluation is a set of vectors $\Theta_l i$ for each text. Now we compare volumes of "boxes" made with these vectors, the volume being considered as the main definitive factor in similarity analysis:

$$V_{P_i} = \prod_{l=1}^{4} \Theta_{li} \tag{3}$$

For each text in the test set, as many box volumes are calcualted as there are classes. After that within each class, the texts are sorted in the decreasing order by these volumes. The bigger the volume is, the more likely it is that the text belongs to this class. Hence, the texts at the beginnig of the list are supposedly relevant. However, we would want to establish a borderline after which we are not likely to meet relevant texts anymore.

3.3 Distribution Function

The Distribution Function states which texts are relevant for the query based on their weight distribution. Note that attribution of a text to more than one class is possible.

Let us first consider weighting a list of texts based on two model texts from the class "football fans" with the help of the Relevance Function. Figure 1 demonstrates a list of 4,030 text weights ("box volumes") sorted in the decreasing order.

It forms an exponent-like curve. The few texts in the left part of it have very high results (these are mainly texts of football communities) compared to the long "tail" on the right. The tail commences after a very steep passage between relevant and non-relevant texts. Hence, the point that separates relevant texts from irrelevant (the break point) should be somewhere at this steep part of the curve. To calculate it, we will analyze difference between weights by the slope of a characterisitc line connecting each point $(x_i; y_i)$ and the X-axis at $(x_i + 1; 0)$.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

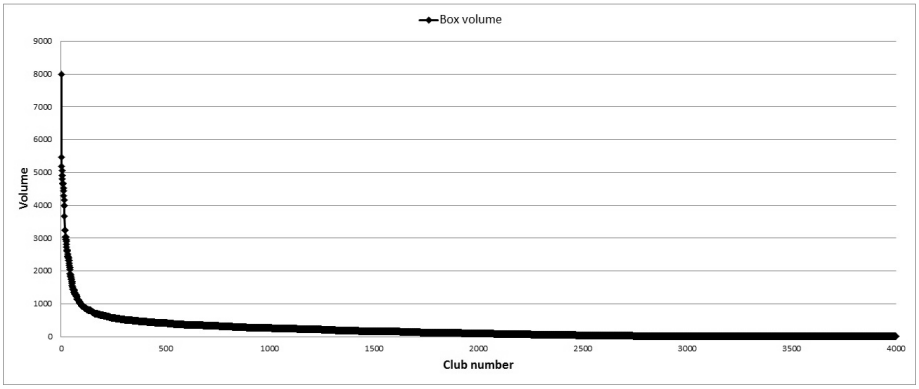

Fig. 1. Box volumes of 4,030 texts evaluated for the MaI "football".

To compare slopes of BC and DE, let us rearrange the diagram so that every segment starts at the point $(x_0; y_0)$ and goes to $(x_n; y_j)$. See Figure 2, on the left.

The slope $a \in [0; +\infty]$ is calculated at the point $(x_1; y_1)$, where $y_i = a \cdot x_1 + b$. As the segment begins at 0, $b = 0$. We calculate x as an arithmetic mean of the text weights:

$$x_1 = \frac{\sum V_{P_i}}{N} \tag{4}$$

So:

$$y_i = a_i \cdot x_1 \Longrightarrow a_i = \frac{y_i N}{V_{P_i}} \tag{5}$$

Empirically, we found out that the best results have $a > 7.01$. Table 2 demonstrates relevant results of the mentioned calculations for the class "football fans".

3.4 Tests

To test the efficiency of our algorithm, we tried several existing implementations of supervised learning algorithms from the "Scikit-learn" package [26] with different optimization parameters: SVM, Neural Networks, Naive Bayes, Logistic Regression, Decision Trees, and k-Nearest Neighbors. The training set included two Model Texts in each of the three classes; the training set for the "Miscellaneous" class was formed with the four Model Texts, belonging to two other classes. For example, for the class of "football fans", two Model Texts go to the training set as class representatives, and the four Model Texts of historical reenactors and vegetarians form the training set for the class "Miscellaneous" [11].

[11] These are conditions similar to what our algorithm requires. To extract features, it needs one or two Model Texts and a couple of non-class texts to extract *uniques*.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

Table 2. Results of the Distribution Function in the class of football fans.

Text rating	Class	Box volume	Slope
1	Football	8000	38.92
2	Football	5460	26.56
3	Football	5187	25.23
4	Football	5054	24.59
5	Football	4921	23.94
6	Football	4921	23.94
7	Football	4800	23.35
8	Football	4680	22.77
9	Football	4662	22.68
10	Football	4662	22.68
11	Football	4536	22.07
12	Football	4446	21.63
13	Misc.	4284	20.84
14	Football	4165	20.26
15	Football	4000	19.46
16	Football	3996	19.44
17	Football	3675	17.88
18	Football	3240	15.76
19	Football	3240	15.76
20	Misc.	3240	15.76
21	Football	3060	14.89
22	Football	3038	14.78
23	Football	2964	14.42
24	Misc.	2940	14.30
25	Misc.	2890	14.06
26	Football	2805	13.65
27	Misc.	2720	13.23
28	Misc.	2640	12.84
29	Misc.	2625	12.77
30	Misc.	2592	12.61
31	Misc.	2520	12.26
32	Football	2448	11.91
33	Misc.	2448	11.91
34	Misc.	2436	11.85
35	Misc.	2400	11.68
36	Football	2380	11.58
37	Misc.	2325	11.31
38	Misc.	2325	11.31
39	Misc.	2240	10.90
40	Misc.	2176	10.59

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

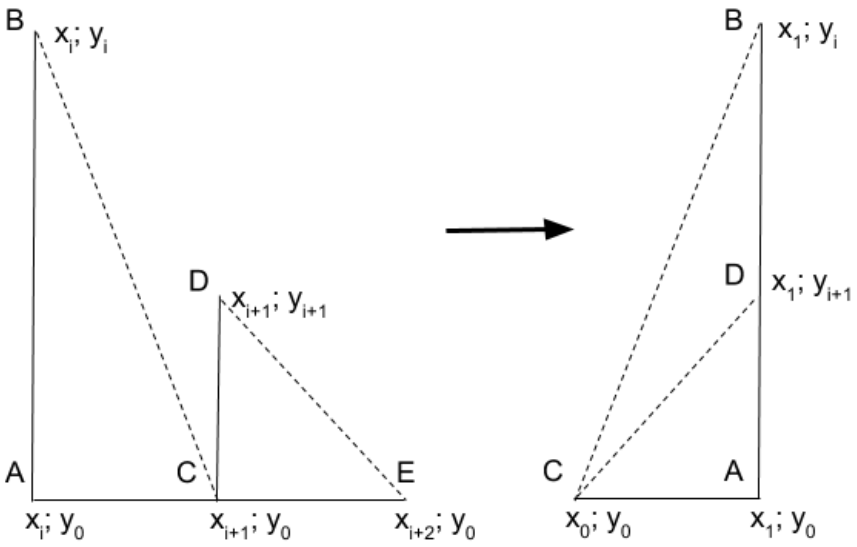

Fig. 2. The slope of the characteristic line.

The test set contained 30 texts of the studied class (e.g. football fans), 30 texts of the two other classes from the training set (e.g. historical reenactors and vegetarians) and 4,000 texts of the class "Miscellaneous" (i.e. not belonging to any of the three). The only algorithm providing a comparable result in such conditions was SVM (with the linear kernel, C=5). Table 3 demonstrates it.

It is of interest that in all the three classes the F-score of our algorithm was very close in value. "Vegetarianism"appears to be the most well-balanced class by the three measures varying within the scale of 0.02. The results would be better if the value of the slope at the break point were optimized for every particular class. But that is the drawback of having just one Model Text without a large set of labeled data. How the break point moves in different classes and with sets of different size is yet an issue to be studied.

4 Conclusions

In the present article, we attempted to describe a new approach to classification of social network pages by interests of users. We suggested that retrieval of pages of interest should be based on one or two Model Texts rather than on a large collection. Even such a classifier as SVM that is typically used with large datasets gives a reasonably good (beyond the chance) classification result with only six texts in the training set and 4,090 texts in the test set. However, we suggested our own supervised learning algorithm that outperforms SVM in the

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

Table 3. Retrieval of the three MaIs from a collection of 4,090 texts.

MaI	Measure	Own algorithm	SVM (linear)
Vegetarianism	Precision	0.73	0.62
	Recall	0.71	0.50
	F1-score	0.72	0.56
Historical reenactment	Precision	0.53	0.18
	Recall	1.00	0.30
	F1-score	0.70	0.23
Football	Precision	0.97	0.52
	Recall	0.52	0.87
	F1-score	0.67	0.65

same conditions. The algorithm can be applied in a recommender system for recommendation of pages of interest based on a page that a user already follows.

In a way, our algorithm can be viewed as a simplified and more intuitive and expertise-based version of SVM, designed for a particular task. It also separates vectors in a hyperspace but in a "fuzzy" way so that one text can be attributed to several classes. However, with the lack of a large set of labeled data for training we cannot be sure that the break point is always the same. In a real life situation, a user can be offered the whole rated list of pages starting with the top results until they stop scrolling for further pages.

As for the further research, we are planning to modify our algorithm for tasks like learning individual user interests and their specification, i.e. when a major interest can be specified into smaller ones which attract subgroups of users. For example, vegetarians call themselves "vegans", "rawatarians", "fruitarians"; football fans support one particular football team; historical reenactors deal with particular periods of time and certain cultures. Finally, we think that detecting a social group automatically when nothing is known about it yet (unsupervised learning of interests) is the most challenging task.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

Bibliography

[1] Agichtein, E., Brill, E., Dumais, S., Ragno, R.: Learning user interaction models for predicting web search result preferences. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3–10. ACM (2006)

[2] Ahmed, A., Low, Y., Aly, M., Josifovski, V., Smola, A.J.: Scalable distributed inference of dynamic user interests for behavioral targeting. In: KDD (2011)

[3] Al-Kouz, A., Albayrak, S.: An interests discovery approach in social networks based on semantically enriched graphs. In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. pp. 1272–1277. IEEE (2012)

[4] Bakalov, F., König-Ries, B., Nauerz, A., Welsch, M.: A hybrid approach to identifying user interests in web portals. In: IICS. pp. 123–134 (2009)

[5] Bentley, A.F.: The process of government. Ripol Klassik (1955)

[6] Billsus, D., Pazzani, M.J.: A hybrid user model for news story classification. In: UM99 User Modeling, pp. 99–108. Springer (1999)

[7] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)

[8] Bonhard, P., Sasse, M.A.: Knowing me, knowing you–Using profiles and social networking to improve recommender systems. BT Technology Journal **24**(3), 84–98 (2006)

[9] Brown, J., Broderick, A.J., Lee, N.: Word of mouth communication within online communities: Conceptualizing the online social network. Journal of Interactive Marketing **21**(3), 2–20 (2007)

[10] Dugan, C., Muller, M., Millen, D.R., Geyer, W., Brownholtz, B., Moore, M.: The dogear game: a social bookmark recommender system. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work. pp. 387–390. ACM (2007)

[11] Firan, C.S., Nejdl, W., Paiu, R.: The benefit of using tag-based profiles. In: Web Conference, 2007. LA-WEB 2007. Latin American. pp. 32–41. IEEE (2007)

[12] Fire, M., Puzis, R.: Organization mining using online social networks. Networks and Spatial Economics **16**(2), 545–578 (2016)

[13] Fischer, G.: User modeling in human–computer interaction. User Modeling and User-adapted Interaction **11**(1), 65–86 (2001)

[14] Frolov, S.: Sociology: personality and society. The main factors of personality development (1994)

[15] Gomaa, W.H., Fahmy, A.A.: A survey of text similarity approaches. International Journal of Computer Applications **68**(13) (2013)

[16] Groh, G., Ehmig, C.: Recommendations in taste related domains: collaborative filtering vs. social filtering. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work. pp. 127–136. ACM (2007)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

[17] Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., Ofek-Koifman, S.: Personalized recommendation of social software items based on social relations. In: Proceedings of the Third ACM Conference on Recommender Systems. pp. 53–60. ACM (2009)
[18] Li, X., Guo, L., Zhao, Y.E.: Tag-based social interest discovery. In: Proceedings of the 17th International Conference on World Wide Web. pp. 675–684. ACM (2008)
[19] Li, Y., Dong, M., Huang, R.: Special interest groups discovery and semantic navigation support within online discussion forums. In: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. pp. 3904–3911. IEEE (2008)
[20] McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. In: IJCAI. vol. 5, pp. 786–791. Citeseer (2005)
[21] Merton, R.K.: Social structure and anomie. American Sociological Review **3**(5), 672–682 (1938)
[22] Mikhalkova, E., Karyakin, Y., Ganzherli, N.: A comparative analysis of social network pages by interests of their followers. arXiv preprint arXiv:1707.05481v2 (2017)
[23] Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69**(2), 026113 (2004)
[24] Pazzani, M.J.: A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review **13**(5-6), 393–408 (1999)
[25] Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: The Adaptive Web, pp. 325–341. Springer (2007)
[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
[27] Piao, G., Breslin, J.G.: Interest representation, enrichment, dynamics, and propagation: A study of the synergetic effect of different user modeling dimensions for personalized recommendations on twitter. In: European Knowledge Acquisition Workshop. pp. 496–510. Springer (2016)
[28] Piao, S., Whittle, J.: A feasibility study on extracting Twitter users' interests using NLP tools for serendipitous connections. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. pp. 910–915. IEEE (2011)
[29] Ramage, D., Dumais, S.T., Liebling, D.J.: Characterizing microblogs with topic models. ICWSM **10**(1), 16 (2010)
[30] Reicher, S.: The determination of collective behaviour. Social Identity and Intergroup Relations pp. 41–83 (1982)
[31] Scott, J.: Social network analysis. SAGE Publications (2017)
[32] Sen, S., Vig, J., Riedl, J.: Tagommenders: connecting users to items through tags. In: Proceedings of the 18th International Conference on World wide web. pp. 671–680. ACM (2009)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18

[33] Shen, W., Wang, J., Luo, P., Wang, M.: Linking named entities in tweets with knowledge base via user interest modeling. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 68–76. ACM (2013)
[34] Shi, L.l., Liu, L., Wu, Y., Jiang, L., Hardy, J.: Event detection and user interest discovering in social media data streams. IEEE Access (2017)
[35] Stefani, A., Strapparava, C.: Exploiting NLP techniques to build user model for Web sites: the use of WordNet in SiteIF Project. In: Proc. 2nd Workshop on Adaptive Systems and User Modeling on the WWW (1999)
[36] Szomszor, M., Alani, H., Cantador, I., OHara, K., Shadbolt, N.: Semantic modelling of user interests based on cross-folksonomy analysis. In: International Semantic Web Conference. pp. 632–648. Springer (2008)
[37] Volkova, S., Coppersmith, G., Van Durme, B.: Inferring user political preferences from streaming communications. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 186–196 (2014)
[38] Wang, Q., Xu, J., Li, H.: User message model: A new approach to scalable user modeling on microblog. In: Asia Information Retrieval Symposium. pp. 209–220. Springer (2014)
[39] Xu, S., Shi, Q., Qiao, X., Zhu, L., Zhang, H., Jung, H., Lee, S., Choi, S.P.: A dynamic users interest discovery model with distributed inference algorithm. International Journal of Distributed Sensor Networks **10**(4), 280–892 (2014)
[40] Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems **42**(1), 181–213 (2015)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_18