BUILDING GRAPH MODELS OF DEPTH ANALYSIS OF DATA OF CLOUD SERVICES SPECIALIZED IN WORKING WITH SOCIAL NETWORKS

Makarov A.E.[1] (Russian Federation), Varlamov A.A.[2] (United States of America)
Email: Makarov576@scientifictext.ru

[1]Makarov Anatoly Evgenievich - Solutions Architect,
IBM,
MOSCOW;
[2]Varlamov Aleksandr Aleksandrovich – Senior Solutions Architect,
LI9 INC,
NORTH CAROLINA, Raleigh, UNITED STATES OF AMERICA

Abstract: *the features of the analysis methods of big data arrays generated by social networks' cloud services are considered. Classification of typical tasks for this class of hardware and software platforms by the generated data format and the analysis system development goals is proposed. In contrast, the urgency of the task of automatic keyword recognition is indicated. To solve this problem, within the framework of the study, a complex methodology for constructing graph models of deep analysis was proposed, which is based on calculating the term frequency, determining the centrality measure, and the position of the nodes in the network. As a result of this work, a technique for recognizing keywords with several attributes was developed.*
Keywords: *graph model, deep analysis, social network, multi-attribute keyword extraction, centrality measures.*

ПОСТРОЕНИЕ ГРАФОВЫХ МОДЕЛЕЙ ГЛУБИННОГО АНАЛИЗА ДАННЫХ ОБЛАЧНЫХ СЕРВИСОВ, СПЕЦИАЛИЗИРУЮЩИХСЯ НА РАБОТЕ С СОЦИАЛЬНЫМИ СЕТЯМИ

Макаров А.Е.[1] (Российская Федерация), Варламов А.А.[2] (Соединенные Штаты Америки)

[1]Макаров Анатолий Евгеньевич - архитектор решений,
IBM,
г. Москва;
[2]Варламов Александр Александрович – главный архитектор решений,
Li9 INC,
Северная Каролина, г. Райли, Соединенные Штаты Америки

Аннотация: *рассмотрены особенности анализа больших массивов данных, которые генерируются в рамках работы облачных сервисов, специализирующихся на работе с социальными сетями. Предложена классификация задач, характерных для данного класса аппаратно-программных платформ в соответствии с форматом генерируемых данных, а также целей, которые преследует разработка системы анализа; при этом указана актуальность задачи автоматического распознавания ключевых слов. Для решения данной задачи в рамках исследования была предложена комплексная методология построения графовых моделей глубинного анализа, которая базируется на расчете частоты появления терминов, определении показателя центральности и положения узлов. В результате проведенной работы была разработана методика распознавания ключевых слов с несколькими атрибутами.*
Ключевые слова: *графовая модель, глубинный анализ, социальная сеть, распознавание ключевых слов с несколькими атрибутами, показатель центральности.*

Введение

Одна из ключевых тенденций развития современных социальных сетей состоит в сокращении объема текстовых блоков при дополнении их массивами мультимедийных данных. Это в значительной степени актуализирует задачу корректного выделения системой автоматического анализа в небольшом объеме текста ключевых слов. Ключевое слово, как функциональный элемент системы анализа, представляет собой одно слово или словосочетание, определяющее контекст. От точности распознавания ключевых слов в значительной мере зависит эффективность работы базовых алгоритмов аппаратно-программной платформы социальной сети, и, в том числе, коммерческий эффект ее эксплуатации. На уровне рассмотрения процедуры глубинного анализа как более высокого уровня абстрагирования также может быть полезен переход к понятию внешнего мема (external meme, e-meme) как логической единицы на базе определенных ключевых слов, которое способно внести изменения в базу знаний, через активацию или замещение внутренних мемов (internal meme, i-meme). Для определения полного набора e-meme необходимо построить адекватную модель текстового графа, узлы которого соответствуют текстовым элементам, а ребра — кратчайшим расстояниям между их позициями в текстовом блоке [1-3].

Следует отметить, что построение модели текстового графа в соответствии с данным подходом подразумевает анализ таких показателей как частота появления терминов, показатель центральности в соответствии с типом центральности, критическое расстояние и топология графа. При построении указанной модели необходимо помнить, что текстовый граф в ряде случаев является несвязным и, соответственно, показатель центральности степени близости, как и показатель центральности по эксцентричности (eccentricity centrality) не всегда может быть задействован. Универсальным подходом при решении этой задачи является применение методики распознавания ключевых слов с несколькими атрибутами [4, 5], что определяет ***актуальность*** данного исследования.

Анализ последних исследований и публикаций в данной области указал на преимущества применения методики распознавания ключевых слов с несколькими атрибутами при построении универсальной модели текстового графа для дальнейшего определения e-meme текстового блока [4, 5]. В первую очередь были рассмотрены методики оптимального выбора меры центральности, что позволило выделить роль узлов в общей архитектуре сети в зависимости от ее типа [6, 7]. Высокая эффективность методов, которые базируются на показателе промежуточной центральности (betweenness centrality, BWC) привела к необходимости рассмотрения исследований [8, 9], где данный подход был реализован путем внедрения моделей распространения (diffusion models) либо построения индикатора положения актора в сети. Также проведенный анализ включал в себя исследование структурных свойств графа через рассмотрение его топологии [10], с учетом того, что корреляции топологических метрик в значительной степени зависит от типа графа. Сложный граф характеризуется топологическими особенностями, которые позволяют определить его связность и предсказать характер процессов, которые выполняются в рамках сети. Для многослойной сети при этом следует использовать как топологические метрики, так и метрики сервисов (service metrics). Кроме того, для решения поставленной задачи, может быть предложен алгоритм по выделению значимых пересечений графов сети [11, 12], в частности анализ значимости пересечений текстовых графов социальной сети на базе методики определения совместной цели (collaborative purpose methods). При таком подходе также строится система индикаторов, которым присваиваются весовые коэффициенты, что позволяет последовательно проанализировать все узлы сети [13].

Представленные методики могут быть рассмотрены как инструментальная база для построения комплексной методологии по работе с моделями текстовых графов и дальнейшего эффективного распознавания в текстовых блоках ключевых слов с несколькими атрибутами, что на данном этапе было предложено выделить как ***нерешенную часть общей проблемы***.

Целью работы, таким образом, стала разработка системы глубинного анализа тестовых блоков и построения соответствующих моделей текстовых графов, которые базируются на определении частоты появления терминов, а также определении показателя центральности и положения узлов.

1. Общие принципы анализа текстовых блоков социальной сети на основе текстового графа

Базовый алгоритм, который позволяет обобщить современные методики по анализу данных текстовых блоков социальной сети (включая глубинный анализ и распознавание ключевых слов с несколькими атрибутами) представлен на рис. 1.

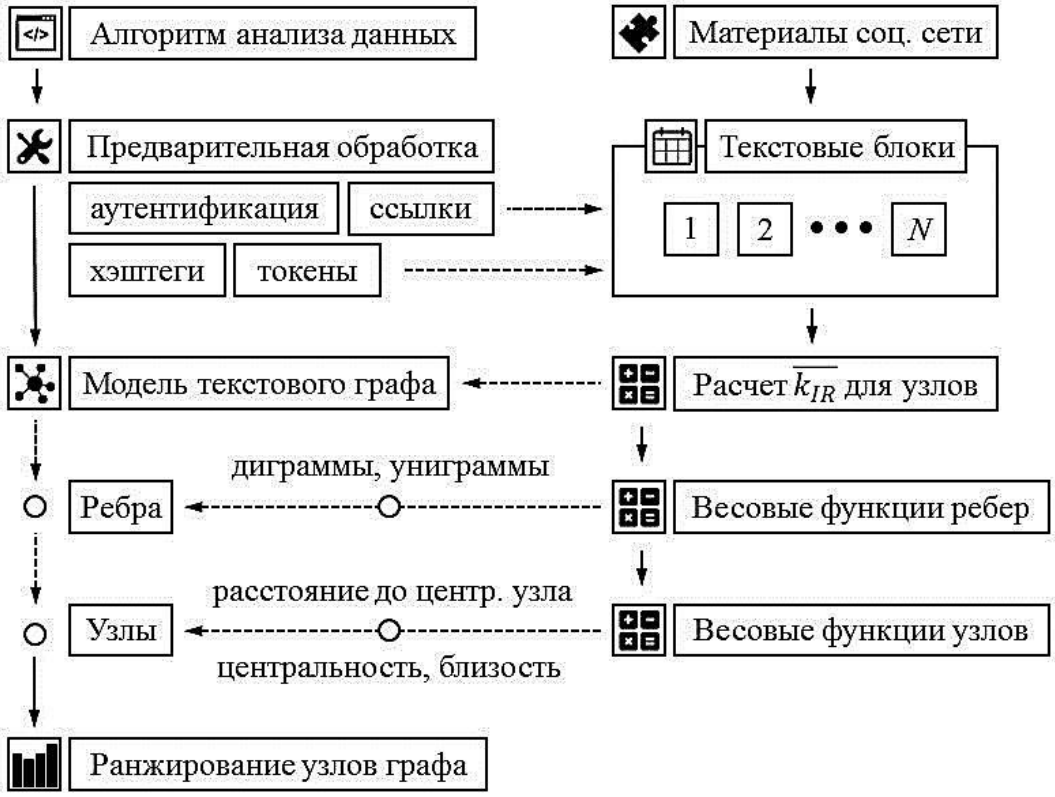

Рис. 1. Базовая схема анализа текстовых блоков социальной сети через построение текстового графа

Как показано на схеме, стандартный подход по построению модели текстового графа подразумевает выполнение процедуры предварительной обработки текстового блока с удалением данных аутентификации автора, ссылок, хештегов, токенов и прочих служебных знаков разметки текста. Работа с массивом обработанных данных включает в себя расчет следующих параметров:

- среднее значение регулярности расположения узлов $\overline{k_{IR}}$ в сети (incidence regularity, IR), которое рассчитывается как соотношение суммы регулярностей узлов заданного типа к общему количеству узлов, и определяет узлы, которые могут быть исключены как не значимые для анализа;
- весовые функции множества ребер графа, на основе которых могут быть построены *N*-граммы (для большинства алгоритмов анализа данных социальных сетей это униграммы и диграммы);
- весовые функции множества ребер графа, на основе которых могут быть определены расстояния узлов до центрального узла, а также коэффициент центральности и степень близости.

Конечным этапом при этом является выполнение процедуры ранжирования узлов и ребер построенного текстового графа.

2. Построение модели текстового графа для распознавания ключевых слов с несколькими атрибутами

Построение текстового графа после выделения начального набора ключевых слов происходит следующим образом:

- определение в качестве вершин графа ключевых слов;
- определение в качестве ребер графа пар ключевых слов;
- расчет коэффициента регулярности для ребер.

Коэффициент регулярности ребра при этом рассчитывается через коэффициенты регулярности его вершин. Определим в рамках построения алгоритма выбора вершин (vertex assignment) ребро $e_{ab} = (a, b)$, где a и b — инцидентные ему вершины, для которых может быть определена регулярность (значения $k_R(a)$, $k_R(b)$ и $k_R(a, b)$). В таком случае регулярность ребра c рассчитывается как:

$$k_R(e_{ab}) = \frac{k_R(a, b)}{k_R(a) + k_R(a) - k_R(a, b)} \quad (1)$$

Далее производится расчет показателя нагрузки на вершину (vertex load), что является основным этапом в определении значимости ключевого слова. Весовой коэффициент узла a сети, где центральный узел определяется как c, может быть определен как функция от следующих аргументов (рис. 2):

- расстояние до центрального узла как величина обратная критическому интервалу $d_C(a) = 1/D(c, a)$;

- коэффициент избирательности узла (selectivity correspondence, SC);
- коэффициент исходящих связей узла (out degree, OD);
- степень близости узла (closeness correspondence, CC);

Рис. 2. Построение модели распознавания ключевых слов в текстовом блоке на базе процедура ранжирования узлов

Коэффициент избирательности узла a рассчитывается на основе соотношения значений коэффициента устойчивости узла (node durability, $k_{\mathrm{du}}(a)$), что определяется как сумма весовых $w(e_{a-i})$ коэффициентов ребер e_{a-i} ($i \in [1; I]$), в отношении которых данный узел является инцидентным, и степени узла (node degree, $k_{\mathrm{de}}(a)$):

$$\begin{cases} k_{\mathrm{SC}}(a) = \dfrac{k_{\mathrm{du}}(a)}{k_{\mathrm{de}}(a)} \\ k_{\mathrm{du}}(a) = \sum\limits_{i=1}^{I} w(e_{a-i}) \end{cases}. \quad (2)$$

Коэффициент исходящих связей узла позволяет определить узлы, которые взаимодействуют с большинством других узлов сети, что указывает на значимость соответствующих им ключевых слов. Данный коэффициент рассчитывается как соотношение соседних узлов N_a к узлу a по отношению к общему числу узлов N помимо того, для которого определяется данный коэффициент:

$$k_{\mathrm{OD}}(a) = N_a/(N-1). \quad (3)$$

Степень близости узла рассчитывается через кратчайшее расстояние между данным узлом и другими узлами $d_{min}(a, a_j)$, где a_j — все соседние узлы по отношению к узлу a ($j \in [1; J]$),):

$$k_{\mathrm{CC}}(a) = \frac{N-1}{\sum_{j=1}^{J} d_{min}(a, a_j)}. \quad (4)$$

Весовой коэффициент узла может быть рассчитан как сумма приведенных выше коэффициентов и расстояния до центрального узла:

$$w(a) = k_{\mathrm{SC}}(a) + k_{\mathrm{OD}}(a) + k_{\mathrm{CC}}(a) + d_{\mathrm{C}}(a). \quad (5)$$

Процедура ранжирования узлов и ребер (node edge ranking, NE-ранжирование) на математическом уровне определяется через функцию $r(a_i)$. Для расчета ее значения необходимо определить весовые коэффициенты w_{ij} ребер e_{ij} инцидентными узлами которых являются узлы a_i и a_j. Соотнесение функций ранжирования $r(a)$ узлов a_i и a_{i-m} может быть рассчитано как:

$$r(a_i) + (k_{\mathrm{D}} - 1) \cdot w(a_i) = k_{\mathrm{D}} \cdot w(a_i) \cdot \sum_{a_j = a_{i-m}}^{a_i} \left(\frac{w(e_{ij}) \cdot r(a_j)}{\sum_{a_k}^{a_j} w(e_{jk})} \right), \quad (6)$$

где k_{D} — коэффициент затухания (damping factor), который определяется на этапе построения алгоритма анализа.

3. Методика оценки эффективности распознавания ключевых слов в текстовых блоках социальных сетей

Для построения системы оценки алгоритма для распознавания ключевых слов на обучающей выборке текстового фрагмента предлагается использовать следующие стандартные показатели (рис. 3):

- точность (accuracy, κ_a), как соотношение числа истинных результатов тестирования и общего числа результатов;
- прецизионность (precision, κ_p), как соотношение числа истинно положительных результатов тестирования и общего числа положительных результатов тестирования;
- полнота (recall, κ_r), как соотношение числа истинно положительных результатов тестирования и суммы истинно положительных и ложноотрицательных результатов тестирования.

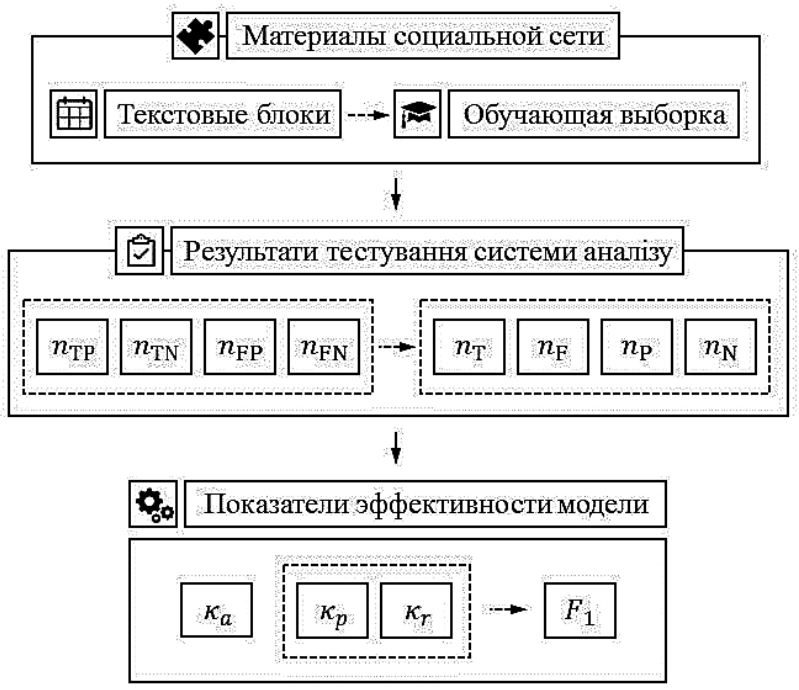

Рис. 3. Алгоритм оценки эффективности распознавания ключевых слов в текстовых блоках

Для их расчета, таким образом, необходимо ввести показатели количества истинно положительных (true positive, n_{TP}), истинно отрицательных (true negative, n_{TN}), ложноположительных (false positive, n_{FP}) и ложноотрицательных (false negative, n_{FN}) результатов тестирования. При этом можно также ввести показатель общего числе результатов $n_{\Sigma} = n_{TP} + n_{TN} + n_{FP} + n_{FN}$, а кроме того общего числа истинных результатов $n_T = n_{TP} + n_{TN}$ и позитивных результатов $n_P = n_{TP} + n_{FP}$.

Представленный подход позволяет помимо точности модели $\kappa_a = n_T/n$ рассчитать также и показатель *F*-меры:

$$F_1 = 2 \cdot \frac{\kappa_p \cdot \kappa_r}{\kappa_p + \kappa_r}, \text{где} \begin{cases} \kappa_p = \frac{n_{TP}}{n_P} \\ \kappa_r = \frac{n_{TP}}{n_{TP} + n_{FN}} \end{cases} . \quad (7)$$

Показатель *F*-меры дает возможность оценить гармоническое среднее между прецизионностью и полнотой модели, а соответственно построить модель, параметры которой будут оптимальным образом сбалансированы.

Выводы

В результате проведенного исследования были рассмотрены методы построения алгоритмов анализа больших массивов текстовых данных, которые генерируются пользователями социальных сетей. Предложена обобщенная схема анализа текстовых блоков через построение текстового графа. Разработана модель распознавания ключевых слов в текстовом блоке, которая базируется на процедурах определения весовых коэффициентов и ранжирования узлов. Для алгоритмов, построенных на основе данного подхода, предложена система оценки, в которой используются показатели точности, прецизионности и полноты, а также комплексный показатель *F*-меры.

Список литературы / References

1. *Shabunina E. & Pasi G.* (2018). A graph-based approach to ememes identification and tracking in Social Media streams. Knowledge-Based Systems, 139, 108-118. doi:10.1016/j.knosys.2017.10.013.
2. *Bordoloi M. & Biswas S.K.* (2018). Keyword extraction from micro-blogs using collective weight. Social Network Analysis and Mining, 8 (1). doi:10.1007/s13278-018-0536-8
3. *Weiler A., Grossniklaus M. & Scholl M.H.* (2016). Editorial: Survey and Experimental Analysis of Event Detection Techniques for Twitter. The Computer Journal. 60 (3), 329–346 doi:10.1093/comjnl/bxw056.
4. *Li Y., Li M. & Shen Y.* (2016). A Multi-attribute Keyword Retrieval Mechanism for Encrypted Cloud Data. International Journal of Security and Its Applications, 10 (12), 335-346. doi:10.14257/ijsia.2016.10.12.27.
5. *Bondade A. R., Patil P. & Patle G.* (2020). Attribute based Encryption for Improved Multi-Keyword Search in Information Network, 2020. 5th International Conference on Communication and Electronics Systems (ICCES). doi:10.1109/icces48766.2020.9138046.
6. *He X. & Meghanathan N.* (2016). Alternatives to Betweenness Centrality: A Measure of Correlation Coefficient. Computer Science & Information Technology (CS & IT). doi:10.5121/csit.2016.61301.
7. *Nikolaev A.R., Giannini M., Meghanathan R.N. & Leeuwen C.V.,* 2019. Enhanced information processing at revisited fixation locations. doi:10.1101/660308.
8. *Benyahia O., Largeron C.:* Centrality for graphs with numerical attributes. In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Pp. 1348–1353. IEEE (2015).
9. *Richiardi J., Achard S., Bullmore E. & Ville D.V.* (2011). Classifying Connectivity Graphs Using Graph and Vertex Attributes, 2011 International Workshop on Pattern Recognition in NeuroImaging. doi:10.1109/prni.2011.18.
10. *Hernández J.M., Van Mieghem P.:* Classification of Graph Metrics, 2011. Pp. 1–20. Delft University of Technology. Mekelweg. The Netherlands.
11. Yang Y., Xie G.: Efficient identification of node importance in social networks. Inf. Process. Manage, 2016. 52 (5), 911–922.
12. *Shi S., Chen K., Wang Y. & Luo R.*, 2011. Node Importance Analysis in Complex Networks Based on Hardware Computing. Journal of Electronics & Information Technology, 33 (10), 2536-2540. doi:10.3724/sp.j.1146.2011. 00363.
13. *Biswas S.K., Bordoloi M., Shreya J.:* A graph based keyword extraction model using collective node weight, 2018. Expert Syst. Appl. 97, 51–59.