DOI 10.33278/SAE-2018.rus.169-172
УДК 004.89
JEL C80

ВИКИ-ГРАФ: АНАЛИЗ СТРУКТУРЫ И ПОТЕНЦИАЛ ПРИМЕНЕНИЯ В ПРИКЛАДНЫХ СЕМАНТИЧЕСКИХ ТЕХНОЛОГИЯХ

СЕРГЕЙ ВЯЧЕСЛАВОВИЧ МАКРУШИН (ORCID 0000-0002-4972-4603)

Финансовый университет при Правительстве Российской Федерации

Аннотация. В настоящее время Википедия - это уже гораздо больше, чем просто самая популярная онлайн-энциклопедия: это уникальный источник частично структурированных знаний о мире. Данные из Википедии активно используются в разных подходах к созданию универсальных семантических сетей. Эти сети являются ядром многих прикладных семантических технологий. В ходе представленного исследования мы создали Вики-граф – сеть русскоязычной Википедии и проанализировали ее структуру. В Вики-графе статьи и категории Википедии рассматриваются как узлы и ссылки в качестве ссылок. Основное внимание в предложенном исследовании уделяется изучению степени распределения узлов.

Ключевые слова: семантические технологии, Википедия, web-граф, теория сложны сетей.

Abstract. Nowadays Wikipedia is much more than the most popular online encyclopedia: it is a unique source of semi-structured knowledge about the world. Data from Wikipedia is actively used in different approaches to create universal semantic networks. This networks are core of many applied semantic technologies. In the research we created the network of the Russian-language segment of Wikipedia and made analysis of its structure. In our network Wikipedia articles and categories are regarded as nodes and references as links. The emphasis of the research is on studying nodes degree distribution.

Keywords: semantic technologies, webgraph, complex network theory.

Определяя семантические технологии в самом широком смысле, можно сказать, что это технологии, которые позволяют оперировать значениями (смыслами) отдельно от данных (например, информационных файлов), а также отдельно от программного кода приложений. С помощью семантических технологий компьютеры получают возможность приблизиться к человеческим возможностям понимания смысла информации, построения логических заключений и обмена смысловыми, а не информационными сообщениями. Семантические технологии сконцентрированы на таких вопросах, как:

- определение понятий и тем;
- извлечение смысла из сообщений;
- категоризация и нахождение взаимосвязей между понятиями.

Но для работы автоматизированных систем со смыслами, их нужно обучить «пониманию» этих смыслов, а для этого, в свою очередь, нужно отчуждать знания людей и формировать их в виде, потенциально подходящем для автоматической интерпретации. В случае построения универсальных семантических систем решение этой задачи очень трудоемко, поскольку требует построения огромных формализованных баз знаний по всем ключевым областям человеческого опыта. Однако для решения этой задачи можно использовать Википедию как ключевой глобальный информационный артефакт частично структурированного знания.

Википедия может являться базой для построения прикладных семантических систем, способных работать с большим количеством понятий реального мира. В связи с этим понимание специфики устройства и построения Википедии как крупнейшего информационного артефакта современной эпохи, обладающего уникальным сочетанием полноты информации о мире и структурированности имеет большую практическую ценность для построения прикладных семантических информаци-

онных систем, в том числе, готовых для использования в бизнес-приложениях.

С технологической точки зрения одним из наиболее удобных способов оперирования информацией, хранящейся в Википедии, является Вики-граф – ориентированный граф, построенный на основе страниц и гипертекстовых ссылок Википедии (соответственно, представляемых как узлы и дуги Википедии). Математический аппарат для анализа Вики-графа предоставляет теория сложных сетей (ТСС) [1] – активно развивающееся с начала двухтысячных годов научное направление, направленное на исследования наблюдаемых эмпирически больших графов.

Объем Википедии достаточно велик для получения очень большого графа (количество документов в Википедии для крупнейших языковых разделов составляет порядка нескольких миллионов документов), но не является чрезвычайно большим и требующим использования очень ресурсоемких ИТ-решений, как, например, веб-граф (множество всех веб-страниц интернета). Кроме того, страницы Википедии создаются членами сообщества по единым правилам, поэтому информация на них (в том числе и ссылки) является частично структурированной, что существенно облегчает ее анализ. На основе и с использованием Википедии можно строить тематические и универсальные семантические сети (онтологии), справочные системы, системы информационного поиска, системы семантического анализа текстов и приложения других типов [2].

Таблица 1

Интегральные показатели сети Википедии, построенной в рамках исследования

Типы страниц	Количество страниц	Ссылок	
		на статьи	на категории
Статьи	3 305 221	92 167 866	9 187 747
Категории	405 363	9 187 747	770 864
Всего	3 710 584	101 355 613	9 958 611

Источник: исследование автора.

В рамках проведенного автором исследования был использован распространяемый фондом Викимедиа (организации, поддерживающей инфраструктуру Википедии) дамп фрагментов базы данных русскоязычного сегмента Википедии. Русскоязычная Википедия насчитывает более 5,3 млн документов, большая часть которых имеет технический характер и более 150 млн ссылок между документами. Для дальнейшего анализа были отобраны только документы из основного пространства имен – типа страниц Википедии, к которому относятся словарные статьи (далее – статьи Википедии) и категории – специализированные документы Википедии, используемые для тематического структурирования информации в Википедии. Как видно из Таблицы 1, категорий и ссылок на них в Википедии примерно на порядок меньше, чем статей и прямых ссылок между статьями.

В среднем на одну статью Википедии приходится 27,9 ссылок на другие статьи и 2,8 ссылок на категории (с помощью таких ссылок статья относится к категории). В отличие от ссылок между статьями ссылки статьи на категорию фактически являются двунаправленными: на всех страницах категорий автоматически формируются ссылки на все статьи, ссылающиеся на эту категорию. В среднем к каждой категории относится 22,7 статей. Наряду с обычными статями категории сами могут быть отнесены к одной или нескольким категориям более высокого уровня,

таким образом в Википедии обеспечивается тематическое структурирование не только статей, но и самих категорий. В среднем каждая категория отнесена к 1,9 категорий более высокого уровня.

Для проведения анализа Википедия была представлена как сеть (граф), узлами которой являются статьи и категории, а ссылками – направленные ссылки между статьями и двунаправленные ссылки для связей с категориями. Анализ был произведен с использованием графовой базы данных neo4j [3]. При загрузке сети из категорий был исключен ряд категорий, имеющих технический характер (около 18 тыс. категорий).

Результат анализа распределения степеней (количества связей) узлов сети, относящихся к статьям Википедии, представлен на рис 1. На графике распределения входящих степеней отражена зависимость количества статей Википедии от количества ссылок на данные статьи. Видно, что в двойном логарифмическом масштабе эта зависимость очень близка к прямой на интервале входящих степеней от 0 до 90, далее наклон прямой незначительно меняется, а зависимость зашумляется и демонстрирует свою дискретную природу для высоких степеней узлов. Вид графика показывает, что входящая степень узлов может хорошо описываться степенным законом распределения $P(k) \sim k^{-\gamma}$ (здесь k – степень узла, а $P(k)$ – доля узлов сети, имеющих степень k (см. рис. 1).

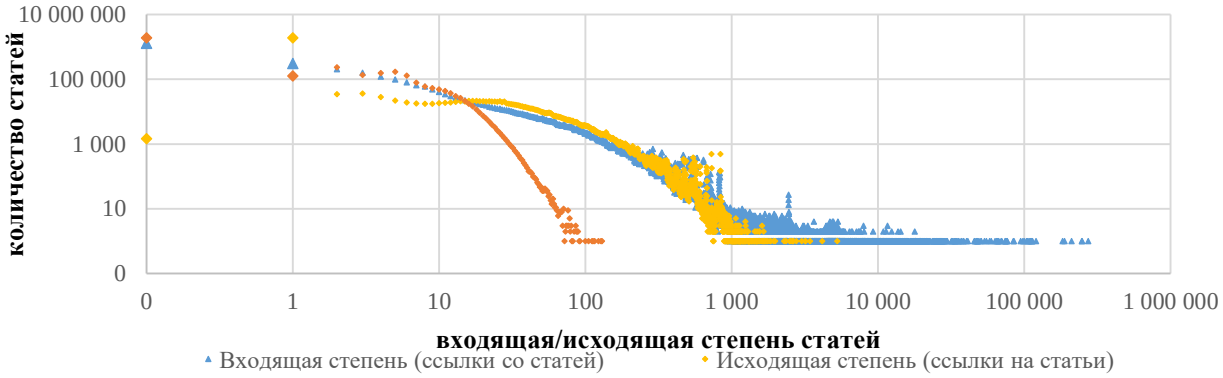

Рис. 1 **Распределение степеней (количества связей) для статей русскоязычной Википедии в двойном логарифмическом масштабе / Log-log plot of nodes degree distribution for article nodes of the Russian-language Wikipedia**

Источник: исследование автора / Source: research of the author

Сети с таким распределением вершин принято называть безмасштабными, их формирование хорошо описывает модель Барбаши-Альберта [4]. Согласно этой модели, сеть возникает в результате пошагового роста сети, при котором действует принцип «предпочтительного присоединения», т.е. новые связи с большей вероятностью образуются у узлов, уже имеющих большое количество связей. Это хорошо согласуется с логикой развития сети Википедии.

В результате роста сети по модели Барбаши-Альберта образуется небольшое количество ключевых узлов сети – «хабов», на распределении степеней узлов это проявляется как «тяжелый хвост» распределения. На рис. 1 видно, что хвост распределения для входящих узлов даже более тяжелый, чем предполагает степенной закон распределения. Аналогичная картина наблюдается для распределения входящих степеней узлов для категорий (см. рис. 2).

Очень существенно от степенного закона распределения отличается распределение степеней узлов для исходящих степеней статей как в голове и середине распределения, так и в хвосте распределения – он не такой протяженный. Это вызвано существенно иной природой возникновения

исходящих ссылок в статьях: они формируются в процессе правки статьи и зависят от ее проработанности и целесообразности добавления новых ссылок, а не от результата большого количества правок на миллионах других статей, поэтому для исходящих ссылок требуется адаптация модели предпочтительного присоединения.

Проведенный в исследовании первичный анализ сети русскоязычного сегмента Википедии показал, что для построения модели формирования входящих ссылок может быть применена модель Барбаши-Альберта, однако для моделирования исходящих ссылок как для статей, так и для категорий потребуется разработка специализированной модели. Сеть Википедии демонстрирует маленькую среднюю длину пути между узлами, и после дополнительного анализа может рассматриваться вопрос ее принадлежности к сетям тесного мира.

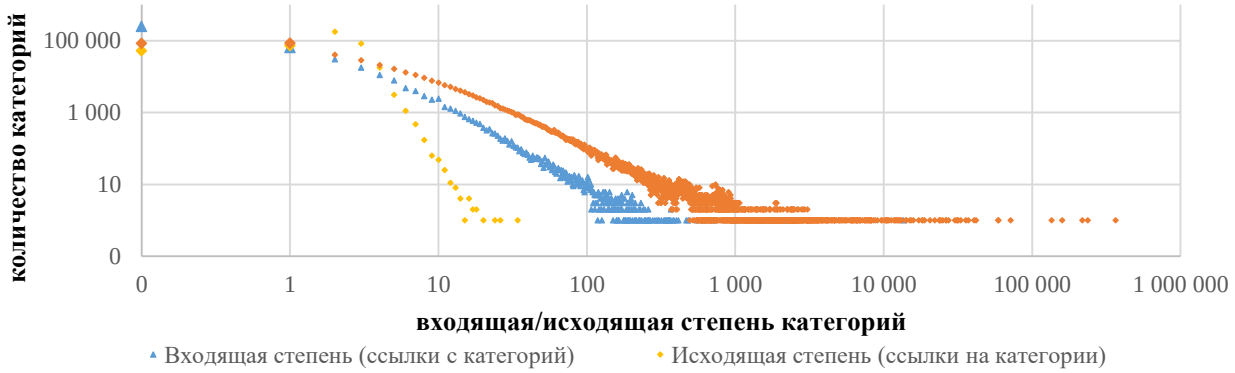

Рис. 2 **Распределение степеней (количества связей) для категорий русскоязычной Википедии в двойном логарифмическом масштабе / Log-log plot of nodes degree distributions for category nodes of the Russian-language Wikipedia**

Источник: исследование автора.

Список источников

1. Евин И.А. Введение с теорию сложных сетей. //Компьютерные исследования и моделирование. 2010, Том 2, N2, с. 121-141
2. М. И. Варламов, А. В. Коршунов Расчет семантической близости концептов на основе кратчайших путей в графе ссылок Википедии //Машинное обучение и анализ данных, 2014. Т. 1, № 8.
3. Neo4j – графовая система управления базами данных с открытым исходным кодом [Электронный ресурс]. – Режим доступа: https://neo4j.com
4. A-L Barabasi, R. Albert, Emergence of scaling in random networks // Science 10/1999 №286 (№5439): 509–512.