Исследование графа категорий английской версии Wikipedia

© А. В. Шкотин
Государственный Геологический Музей РАН
Москва
ashkotin@acm.org

Аннотация

Wikipedia является выдающимся проектом по накоплению знаний, как общего пользования, так и различных областей специализации. Проверка качества этих знаний, особенно автоматическая, чрезвычайна важна. В работе представлены результаты изучения строения английской версии ГКВ (орграф категорных статей Википедии) в целом. Являясь по идее системой тем он поддерживает систематизацию знаний и мы интересуемся из чего эта систематизация состоит и как она устроена. Показано, что в графе есть неприемлемые логические нарушения и обсуждаются организационные и технические методы их устранения.

1 Введение

ГКВ [1] есть подграф графа в котором статьи Википедии приписаны категорным статьям. Выделение ГКВ из этого полного графа есть первая техническая задача. Важно, что далее изучается дамп ГКВ на некоторый момент времени и в нём есть незавершённая "строящаяся" часть. Поэтому выводы надо делать с осторожностью. Естественно ввести термин "точка роста", когда мы натыкаемся "в дампе" на часть, которая ещё не завершена. Дамп полного графа получен из ИСП РАН и соответствует 16 сентября 2010г. Дамп состоит из двух текстовых файлов: файла отображения номера страницы Википедии в номер категорной страницы, что приписывает страницу категории; а также файла в котором номеру страницы Википедии приписано её наименование. Математически ГКВ есть орграф каждый узел которого взаимно-однозначно соответствует категорной странице и помечен её номером. Дуга из узла N1 в узел N2 идёт тогда и только тогда, когда страница с номером N1 есть под-категория страницы с номером N2. Всего таких стрелок 1221133.

В статье исторически вместо термина «дуга» употребляется термин «стрелка».

Множество узлов ГКВ (593796 штук), как и любого произвольного графа, распадается на два подмножества: изолированные узлы (26272) и узлы связанные стрелками (567524 узлов). Изолированная категория это скорее всего "точка роста": на момент снятия дампа она уже есть, но в ГКВ ещё не включена.

Далее анализируется только "граф стрелок", т.е. все характеристики даны без учёта изолированных узлов. Состав изолированных узлов можно посмотреть в отчёте [4] (далее - отчёт) в таблице указанной во введении. Состав и характеристики узлов со стрелками можно посмотреть в таблице указанной там же, равно как и граф стрелок. Важный вопрос - количество связных компонент графа, т.к. в дальнейшем их строение можно изучать отдельно. Таких компонент оказалось 1987. Изолированные узлы при этом учитываются отдельно. Алгоритм разбиения описан в отчёте [5]. Впрочем проще воспользоваться программой, например, Pajek [2] умеющий разбивать узлы графа на слабо связные компоненты.

Первые 10 самых больших компонент:

cn	count	cn	count
1	561636	6680	20
21727	210	19212	19
14332	36	20868	19
2863	29	13325	17
20842	27	13287	16

Здесь cn - уникальный номер компоненты, присвоенный при разбиении. Конечно в случае с Wikipedia малые компоненты это точки роста. Петель (N1 → N1) в графе нет.

Источников (узлов в которые нет входящих стрелок) - 345597. Это категории нижнего уровня. Стоков (узлов из которых нет исходящих стрелок) - 11767. Это категории верхнего уровня дампа и скорее всего "точки роста". Промежуточных узлов, соответственно - 210160.

Максимальное количество исходящих из одного узла стрелок - 85. Это промежуточный узел № 690451, а заголовок "Category:World War II", т.е. эта категория приписана 85 над-категориям. Максимальное количество входящих стрелок (12625)

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

имеет промежуточный узел № 692309 с проясняющим заголовком - "Category:Albums by artist".

2 Анализ заголовков

Заголовки всех узлов категорий (включая изолированные) можно посмотреть в отчёте в таблице указанной в разделе «Анализ заголовков». Таблица содержит - 584606 узлов. Таким образом 9190 узлов ГКВ не имеют заголовков. Они ждут своего исследователя. Анализ текстов заголовков даже безотносительно их подчинения отдельная увлекательная задача. Но начать надо с использованного состава букв.

2.1 Алфавит

Рассмотрим состав букв (characters), употреблённых при именовании категорных статей. Текстовый файл (UTF-8) содержащий состав алфавита можно посмотреть в прикреплении cat2title.abc0.txt к отчёту. Как разделитель букв используется '|'. Вот он:

```
| |!|"|&|'|(|)|*|+|,|-
|./|0|1|2|3|4|5|6|7|8|9|:|;|?|@|A|B|
C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U
|V|W|X|Y|Z|a|b|c|d|e|f|g|h|i|j|k|l|m|
n|o|p|q|r|s|t|u|v|w|x|y|z|~|¡|ª|°|µ|·
|º|½|Á|Â|Ä|Å|Æ|Ç|É|Í|Î|Ñ|Ó|Õ|Ö|×|Ø|Ú|
Ü|Þ|ß|à|á|â|ã|ä|å|æ|ç|è|é|ê|ë|ì|í|î|ï
|ð|ñ|ò|ó|ô|õ|ö|ø|ù|ú|û|ü|ý|þ|ÿ|Ā|ā|ă|
ą|Ć|ć|Č|č|ď|Đ|đ|ē|ė|ę|ě|ğ|Ħ|ħ|Ī|ī|İ|ı
|ļ|Ľ|ľ|Ł|ł|ń|ņ|ň|Ō|ō|ő|œ|ř|Ś|ś|Ş|ş|Š|
š|Ţ|ţ|ť|ū|ŭ|ů|ŵ|ź|Ż|ż|Ž|ž|ơ|ș|ʻ|ḷ|ḻ|ḵ|ṭ|ạ
|ầ|ậ|ắ|ế|ề|ể|ễ|ệ|ị|ọ|ộ|ụ|ỳ|ỹ|–|—
|‘|’|…|共|台|和|國|歲|灣|萬|！|
```

В прикреплении id2title.abc.txt к отчёту можно посмотреть впечатляющее разнообразие букв заголовков всех статей Wikipedia(en).

2.2 Термины в заголовках

Это может быть отдельное важное исследование. Например, количество заголовков в которых встречается слово album - 17591.

3 Стоки

В Приложении-1 к отчёту можно посмотреть начало таблицы стоков с самым большим количеством входящих стрелок.

3.1 От Анастасии к Музыке

В Приложении-2 к отчёту можно посмотреть путь рекордсмен предоставленный Антоном Коршуновым.

Самый длинный путь - 294 вершины. Его начальная категория - № 5760285 Category:Anastacia songs, а конечная - № 691484 Category:Music.

4 Строение ГКВ в целом

В работе [3], с.9 указывается что в ГКВ есть циклы. По идее циклы это аномалия на графе подчинения категорий. И должны занимать малую его часть. Назовём для краткости объединение ор-циклов графа и ор-путей между циклами - *ядро*, а дополнительную часть графа - *мантия*. Стрелки же между мантией и ядром назовём - *связующие*. Таким образом в целом граф состоит из ядра, мантии и связующих стрелок часть из которых идёт из ядра в мантию, а часть - из мантии в ядро. Чтобы выделить ядро, был применён следующий алгоритм:
1. Находим в графе источники и стоки и удаляем их.
2. Если в графе не осталось узлов то стоп.
3. Если есть источники или стоки то идти на 1. Стоп.

5 Ядро ГКВ

5.1 Состав ядра

Количество стрелок в ядре - 38538. Узлов же 13545. Граф ядра опубликован в таблице указанной в разделе «Состав ядра» отчёта. Далее было выполнено "расщепление" ядра на связные компоненты. Это особенно важно, т.к. пути между циклами, сами не входящие в циклы, составляют самостоятельную интересную часть ядра. Оказалось, что имеется одна большая компонента - 13507 узлов. И ещё 19 пар узлов. Характеристики узлов ядра включая разбиение на связные компоненты можно посмотреть в таблице указанной в разделе «Состав ядра» отчёта.

Рассмотрим компоненту №764 ядра. Это пример пары, которая является даже связной компонентой не только ядра, а самого ГКВ. В компоненте два узла:

28736601, Category:Wikipedia sockpuppets of ShantanuSingh198
и
28736686, Category:Suspected Wikipedia sockpuppets of ShantanuSingh19

В Wikipedia они также ссылаются друг на друга и больше ни на что.

Анализ. Что бы ни обозначало "Wikipedia sockpuppets of ShantanuSingh198" очевидно что нечто под него подпадающее (как под понятие) не может быть одновременно лишь "подозреваемым" на подпадание. Равно как и наоборот, т.е. логически эти две категории не пересекаются. И обе стрелки должны быть удалены. Отношения же между ними на, например, OWL 2 [7] должно было бы быть:

```
DisjointClasses(
wcg:Wikipedia_sockpuppets_of_Shantanu
Singh198,
wcg:Suspected_Wikipedia_sockpuppets_o
f_ShantanuSingh198)
```

При этом более правильно ссылаться в обоих статьях друг на друга через тэг Wikipedia «See also».

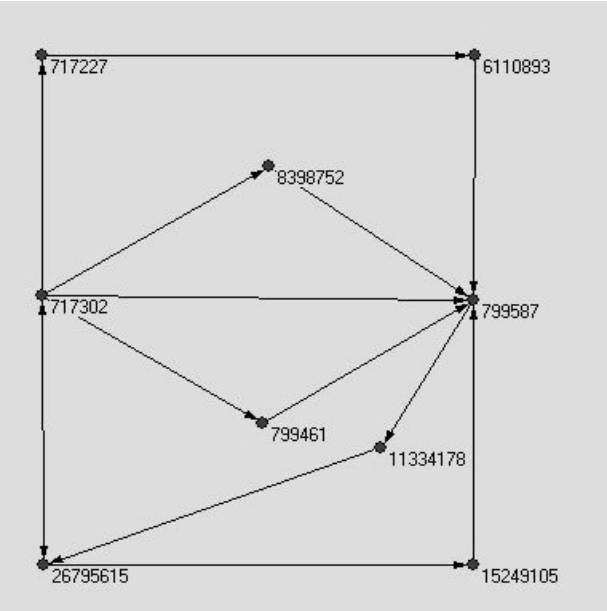

Рис. 1 Рисунок графа компоненты ССК 41

5.2 Сильно связные компоненты ядра

В ядре нас интересует зацикливание отношения под-категория - над-категория. Тут есть два подхода:

- общий - применить алгоритм поиска сильно связных компонент (ССК);

- частный - найти так называемые "линзы" - два узла ссылающиеся друг на друга (как под-категория - над-категория).

Второй путь вполне приемлем для ГКВ, т.к. по идее в нём вообще не должно быть циклов. Впрочем как для линз так и для циклов большей длины следует заметить, что они математически утверждают эквивалентность соответствующих терминов, т.е. синонимию, что в принципе возможно. Но конкретно в Википедия может быть реализовано через redirect. Интуитивно же в большинстве случаев мы обнаружим ошибку, т.е. какие-то стрелки цикла ошибочны.

Чтобы получить состав сильно связных компонент ядра была использована программа Pajek [2]. Заметим, что петель в ГКВ нет, а поэтому узлы ядра не попавшие в ССК это узлы на путях между циклами (см. выше).

ССК оказалось 457. Узлов не входящих в ССК, так сказать связующих ядра — 7646. Есть одна гигантская по сравнению с остальными ССК - в ней 3967 узлов.

В отчёте в разделе «Сильно связные компоненты ядра» приведена таблица самых больших ССК. Рассмотрим для примера компоненту №41 у которой всего 9 узлов (см. рис. 1).

Если номер накладывается на стрелку то под ним наконечника (треугольничка) нет. Это важно т.к. Pajek рисует "линзы" (У1 → У2 → У1) как одну стрелку с наконечниками на обоих окончаниях. В данной ССК линза одна - слева внизу вертикально.

Заголовки узлов:

ni	title
717227	Category:Orthodox rabbis
717302	Category:Talmud rabbis
799461	Category:Mishnah
799587	Category:Talmud
6110893	Category:Talmudists
8398752	Category:Talmud people
11334178	Category:Rabbinic literature
15249105	Category:Talmud concepts and terminology
26795615	Category:Chazal

5.3 Линзы

Линза это два узла такие что: У1 → У2 и У2 → У1. Она может быть отдельной ССК, а может входить в ССК как часть.

В ядре оказалось 1269 линз. Из них 1260 имеют заголовки для обоих узлов. Их можно посмотреть в таблице указанной в разделе «Линзы» отчёта.

6 Мантия - ациклическая часть ГКВ

Чтобы получить мантию мы удаляем из ГКВ ядро. При этом оказывается, что часть источников и стоков станет изолированными. В первом случае все из них исходящие стрелки попали в ядро, во втором - все входящие в них стрелки шли из ядра. Изолировавшихся источников - 14421, а стоков - 60.

Кроме того в мантии появляются ложные вершины (пики). Это те её узлы, которые стали стоками после удаления ядра, а вообще-то имели исходящие стрелки, которые все попадали в ядро. Таких вершин 18157. Причём максимальная высота - 28. Для сравнения, стоков ГКВ получивших уровень, т.е. не изолированных - 11707, максимальная высота - 24.

Ложная вершина - рекордсмен (высоты 28) имеет №15715670, а заголовок "Category:Creation myths".

Замечание. Конечно ГКВ можно представить и в виде "галстука-бабочки" как в работе [6], где орграф был использован для представления схемы связей между транснациональными корпорациями. Но в данном случае сравнение с горами нагляднее - вверх к более обширным темам; горами в которых есть ядро из 20-ти связных компонент. Одна из которых большая, а 19 - линзы.

Количество узлов на уровнях показано ниже в табличке и оправдывает сравнение с горами:

level	count	level	count	level	count
NULL	14481	*NULL*	14481	*NULL*	14481
28	1	18	50	9	1915
27	2	17	57	8	3103
26	3	16	71	7	4858
25	3	15	100	6	7754
24	5	14	149	5	13019
23	7	13	226	4	23302
22	12	12	425	3	45323
21	16	11	697	2	105958
20	20	10	1187	1	331205
19	30			0	13545

В таблице в строке с level = NULL - количество изолированных узлов мантии, а у 0 - количество узлов в ядре.

7 Связующие стрелки

Между мантией и ядром есть стрелки-связующие. Стрелок из ядра в мантию - 591. Стрелок из мантии в ядро — 210514.

8 Другие способы исследования

Можно напрямую изучать http://dbpedia.org через точку входа для SPARQL: http://dbpedia.org/sparql. Привязка к категории идёт через свойство http://purl.org/dc/terms/subject.

Вот пример запроса, который начинает выдавать полный граф связи страниц и категорий:

select ?x ?z where {?x dcterms:subject ?z}

Надо только поставить timeout, например, 1000.

Запрос

```
select ?x ?z where {?x skos:broader ?z}
```

выдаёт отношение "x is a sub-category of z". см. с. p.5 "Categories." [3]

А вот запрос

```
select ?x ?z where
{?x skos:broader ?z. ?z skos:broader ?x.}
```

выдаёт "линзы".

Вот узлы первой:

```
http://dbpedia.org/resource/Category:Poli
tical_philosophers
```
и
```
http://dbpedia.org/resource/Category:Poli
tical_theorists
```

Она действительно есть в Wikipedia(en).

А всего запрос выдаёт 2000 линз, что наверно не предел.

9 Заключение

Естественно считать, что ГКВ должен быть ациклическим графом. Таким образом исследование показало, что аномалии значительны.

Можно создать средства, которые обнаруживая аномалию, например линзу, будут размещать на соответствующих страницах в Discussion уведомление о логическом противоречии.

Основных вопросов два:

1. Как к такому подходу отнесутся авторы страниц категорий? Это можно проверить экспериментально.

2. Как к логическим противоречиям относятся идеологи Википедии? Те кто задаёт правила классификации. Судя по всему индифферентно.

Общая рекомендация. Многие отношения между категориями попавшие в sub-category of следует перенести в See also.

Оценить предстоящую работу можно так: для начала надо разобраться с 1269 линзами. Они сильно убавят размер ССК.

Только если это нужно википедистам можно было бы продолжить и:

- Исследовать длинные пути.
- Попытаться представить архитектуру графа в целом. Например применить 3D визуализацию.
- Проанализировать состав и логику связи заголовков (особенно ССК).

Особняком стоит задача получить и проанализировать русский ГКВ. В проекте dbpedia можно получить дамп русской версии, надо только перекодировать с rdf-кодов букв (например, \u0432) в UTF-8.

Литература

1. Anton Korshunov, Denis Turdakov, Jinguk Jeong, Minho Lee, Changsung Moon. A Category-Driven Approach to Deriving Domain Specific Subset of Wikipedia. Proceedings of SYRCoDIS'11: The Seventh Spring Researchers Colloquium on Databases and Information Systems, 2011, pp. 43-53.
2. Batagelj V., Mrvar A. Pajek reference manual. Ljubljana, April 16, 2012.
3. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. DBpedia - A Crystallization Point for the Web of Data. May 25 2009.
4. Шкотин А., Исследование графа категорий английской версии Wikipedia, Сообщение о результатах первого этапа, Интернет, 2011. https://sites.google.com/site/alex0shkotin/grafy/wikipedia-category-graph
5. Шкотин А., Разбиение графа на связные компоненты, Алгоритм и программа, Интернет, 2011. https://sites.google.com/site/alex0shkotin/grafy/svaznye-komponenty
6. Stefania Vitali, James B. Glattfelder, Stefano Battiston. The network of global corporate control. ArXiv.org, 2011 http://arxiv.org/abs/1107.5728
7. OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax. Boris Motik, Peter F. Patel-Schneider, Bijan Parsia, eds.

W3C Recommendation, 27 October 2009.
http://www.w3.org/TR/owl2-syntax/

Investigation of the English version of the Wikipedia categories graph

Alexander Shkotin

Wikipedia is the outstanding project of knowledge accumulation. The knowledge is both of the general use, as well of various specialization domains. Quality check of this knowledge, especially automatic, is very important. In this paper results of studying of a structure of the English version of WCG (Wikipedia Categories Graph) as a whole are presented. WCG is a system that supports structure of knowledge and we are interested in WCG content and its arrangement. It is shown that in graph there are unacceptable logical violations; organizational and technical methods for their elimination are discussed.