Ассоциативная сеть понятий, образующих запросы к Интернету

И.А. Большаков

Е.И. Большакова

А.Ф. Гельбух

Резюме

В базе пользовательских запросов поисковиков Google и Яндекс выявлена обширная совокупность сочиненных пар существительных.

На их основе построена и описана ассоциативная сеть понятий, из которых часто формируются русскоязычные запросы к Интернету.

Показано, что выявленные пары существительных представительно входят и в текстовые массивы Интернета.

Исследована полученная ассоциативная сеть и составляющие ее понятия.

Задачи данного сообщения

- Описать имеющуюся коллекцию сочиненных именных пар до и после пополнения ее данными из Google и Яндекса;
- Дать приближенную интерпретацию ряда характерных запросов в виде сочиненных пар, показав на примерах несводимость возникающих ассоциаций к семантическим связям типа WordNet
- На основе статистических данных показать, что новые пары понятий встречаются и на сайтах Интернета, а потому могут считаться принадлежащими русскому языку в целом
- Бегло описать созданную из компонентов сочиненных пар ассоциативную сеть понятий, которыми оперирует русскоязычный пользователь в запросах к Интернету
- Проанализировать построенную сеть глубже, выявив понятия с максимальным количеством ассоциативных связей, вычленив и описав связные компоненты сети и др.

Связи внутри сочиненных пар в прежней коллекции пар

- Когипонимы в некой родовидовой иерархии (*руки и ноги, аксиомы и теоремы, труд и капитал, акушерство и гинекология*);
- Синонимы, квазисинонимы и повторы (*траур и скорбь, горести и несчастья, тысячи и тысячи*);
- Антонимы, квазиантонимы, противоположные понятия и конверсивы (*бедные и богатые, актив и пассив*, *Бог и дьявол*, *купля и продажа*, *действие и противодействие*);
- Парные названия и исторически связанные имена (*Босния и Герцеговина, Адам и Ева*).
- *Редко: соучастники некой ситуации (писатель и читатели, закон и порядок, кожа и косметика) или понятия, связанные причинно-следственными связями (война и разруха, преступление и наказание, штормы и наводнения).*

Методика пополнения коллекции

1. Для пар X_i *и* Y_i исходной коллекции (0-й версии) делается попытка найти все новые пары X_i *и ?* и Y_i *и ?* в БДЗ. Этим создается 1-я версия.
2. Для пар X_i *и* Y_i 1-й версии делается попытка найти все новые пары Y_i *и ?* Этим создается 2-я версия.
3. Для пар X_i *и* Y_i 2-й версии делается попытка найти все новые пары Y_i *и ?* Этим создается 3-я версия.

   ……
4. Проверяются и отдельные случаи вхождения в основные массивы Интернета пар «и X_i »

Примеры связей внутри новых сочиненных пар

- Запрос *X и цены* эквивалентен предикату *цены*(*X*)? (Но: *цены и комплектация / наличие / скидки / ценообразование)*
- *При Y = беременность* или *здоровье* запрос представим в виде *влияние*(X,Y)?
- При X или Y = *СМИ* запрос представим симметрич-ным предикатом *взаимодействие*(X,Y)?
- При X = *йога,* Y = *православие / христианство / буддизм имеем* симметричный предикат *совместимость*(*йога*,Y)?
- Пара *ангина и керосин* предполагает структуру с двумя вложенными предикатами: *эффективность*(*лечение*(*ангина*, *керосин*))?

Статистика образцов запросов и ответов

VQ – число запросов, **VS** – число прямых ответов,

VF – число косвенных ответов, все в тысячах

Сочиненная пара	VQ	VS	VF
беременность и роды	1470.0	1380.0	1720.0
беременность и простуда	219.0	249.0	263.0
беременность и компьютер	784.0	99.2	834.0
беременность и месячные	271.0	201.0	251.0
беременность и курение	494.0	52.0	499.0
беременность и питание	1450.0	37.8	1470.0
беременность и грипп	460.0	258.0	593.0
беременность и молочница	171.0	125.0	163.0
здоровье и красота	99700.0	2110.0	144000.0
здоровье и материнство	108.0	118.0	195.0
здоровье и спорт	315000.0	173.0	261000.0
здоровье и комфорт	915.0	178.0	926.0
здоровье и здоровый образ жизни	1960.0	81.6	1180.0
здоровье и долголетие	243000.0	40.1	310000.0
здоровье и окружающая среда	558.0	121.0	426.0

Соотношения статистик в базе данных запросов и в Интернете

Сопоставление векторов статистик велось по известной формуле косинуса

$$COS(VQ,VS) = \frac{\sum_{i}\left(VQ_i VS_i\right)}{\sqrt{\sum_{i}\left(VQ_i\right)^2 \sum_{i}\left(VS_i\right)^2}},$$

что дало

***COS*(*VQ*,*VF*) =** 0,95
- вектора БДЗ и косвенных ответов коллинеарны

***COS*(*VS*,*VQ*) =** 0,26 - вектор прямых ответов идет

***COS*(*VS*,*VF*) =** 0,27 - под углом к векторам БДЗ и косвенных ответов

Наша ассоциативная сеть - это

неориентированный граф с вершинами, помеченными понятиями, входящими в сочиненные пары. Ребра графа соединяют вершины X и Y, если последние образуют сочиненную пару *X и Y* и/или *Y и X*.

Понятия теории графов

- **Степень вершины** это число ребер, которым она принадлежит
- **Висячая вершина** это вершина степени 1
- **Мощностью** графа это число узлов в нем
- **Диаметр** графа это длина самой длинной из кратчайших цепей, связывающих какие-либо две вершины графа
- **Мост** это ребро, разрыв которого увеличивает число связных подграфов
- **Точка сочленения** это вершина, удаление которой ведет к увеличению числа связных подграфов

Примеры вершин сети с их ассоциациями

- *аденоиды:* *аллергия, бассейн, гланды, гомеопатия, кашель, лазеротерапия, миндалины, слух*
- *ангина:* *антибиотики, беременность, гомеопатия, грудное вскармливание, кашель, керосин, мороженое, прополис, сердце, фарингит*
- *аргументация:* *доказательство, контраргументация, опровержение, риторика*
- *аритмия:* *алкоголь, армия, беременность, остеохондроз, роды, спорт, тахикардия*
- *астрономия:* *астрология, астрофизика, космонавтика, космос, непознанное, общество, телескопостроение, физика*
- *безработица:* *бедность, занятость, инфляция, кризис, рынок труда*
- *биотехнология:* *генная инженерия, медицина, микробиология, окружающая среда, селекция, сельское хозяйство, энергетика*

Степени *D* наиболее популярных понятий

D	Понятие	*D*	Понятие	*D*	Понятие
302	*беременность*	36	*право*	27	*власть*
110	*здоровье*	34	*температура*	27	*реклама*
87	*алкоголь*	34	*характер*	27	*экология*
87	*цены*	33	*бизнес*	26	*структура*
54	*спорт*	33	*дизайн*	25	*философия*
52	*культура*	32	*кризис*	24	*контроль*
51	*похудение*	32	*развитие*	24	*наука*
49	*дети*	31	*политика*	24	*пиво*
48	*человек*	31	*ремонт*	24	*христианство*
41	*диабет*	29	*армия*	23	*водка*
40	*диета*	29	*методы*	23	*государство*
39	*курение*	29	*экономика*	23	*деньги*
39	*любовь*	28	*давление*	23	*Интернет*
37	*общество*	28	*лечение*	23	*искусство*
37	*религия*	28	*функции*	23	*православие*
37	*Россия*	27	*безопасность*	23	*прыщи*

Степени *D* популярных многословных понятий

D	Понятие	*D*	Понятие
22	*окружающая среда*	9	*заработная плата*
20	*щитовидная железа*	9	*культура речи*
16	*кормление грудью*	9	*Новый год*
14	*лунный календарь*	9	*общественное мнение*
13	*грудное вскармливание*	9	*социальный контроль*
12	*социальная политика*	8	*бронхиальная астма*
12	*характерные черты*	8	*зеленый чай*
11	*государственное управление*	8	*знаки зодиака*
11	*группа крови*	8	*информационные технологии*
11	*международное право*	8	*образ жизни*
11	*охрана окружающей среды*	8	*оливковое масло*
10	*охрана природы*	8	*охрана труда*
10	*рынок труда*	8	*рыночная экономика*
10	*экономический рост*	8	*социальная справедливость*
9	*витамин С*	8	*тепловые двигатели*
9	*глобальные проблемы*	7	*валютный курс*

Общая характеристика сети (на январь 2010 г.)

- Число понятий в сети 9200
- Суммарное число связанных с ними понятий 25300
- Всего связных подсетей 870
- Доминирующая подсеть включает 56% всех вершин сети
- В доминирующей сети висячие вершины («торчащие иголки») составляют 52%
- Следующая по мощности подсеть в 24 раза меньше доминирующей
- Подсетей из двух вершин 75%
- Среднее число связей у вершины 2,75

Наиболее крупные подсети

Мощ-ность	Под-сетей	Длина диам.	Примеры диаметров	Примеры мостов	Точки сочл.	Тематика
5129	1	14+	***продавцы–покупатели –поставщики–закупки –снабжение– комплектация–цены –ламинат–вода– ветер –снег–грозы– дожди;***	***гололедица–снег; комплектация –цены;***	***цены; снег; вода; водка;***	общежитейский универсум
21	1	10	***любители–профессионалы –дилетанты–специалисты –ЕГЭ–вузы–школы1– колледжи–лицеи–гимназии; ясли–детсады***;	***специалисты–ЕГЭ; лицеи–гимназии;***	***вузы; ЕГЭ; детсады; институты;***	воспитательно-образовательная сфера
13	2	7	***фасад–кровля–фасады– кровли–крыша1– перекрытия–пустоты;***	***изоляция–кровли; кровли–фасады;***	***кровли; крыши;***	(1)детали домов (2)преступность
11	3	6-8	***диаметр–окружность– круг1–крест–шар –сфера1;***	***диаметр– окружность;***	***крест; круг1;***	(1)геометр. фигуры (2)фазы изменения (3)стройматериалы

Некоторые параметры
на май 2010 г.

- Число сочиненных пар 16942
- Из них из существительных 15360
- Число понятий в сети 9700
- Суммарное число связанных с ними понятий 26838
- Среднее число ассоциаций у понятия 2,77

Общие свойства понятий-компонентов ассоциативной сети

- Обычно нейтрального стиля
- В большинстве своем широко используются в обычной речи
- В рамках ассоциаций имеют четко фиксированное значение (как у терминов)
- Однозначно переводятся на иные языки
- Порядка 10% состоят из двух и более слов.
- Если у понятия есть оба числа, но обычно используется множественное

Для чего можно использовать сеть?

- Автоматизированное составление запроса к Интернету в типовых случаях
- Дальнейшие исследования:
  - Сравнение «профиля» русскоязычного пользователя с общемировым
  - Построение антологий для Интернета
  - Сопоставление с другими ассоциативными и идеографическими словарями

Замечания под конец

- Google с 10 марта перестал давать статистику запросов. Это не первый раз, когда гуглисты показывают лингвистам конфетку и почти тут же убирают!
- Развитая нами ассоциативная сеть выложена в Интернете. Если будет интерес, напишите, я выложу самую последнюю версию. При ней дается расшифровка омонимов. Можно выложить и обнаруженные синонимы (их немного).

Спасибо за внимание!
Жду вопросов.

Большаков
Игорь Алексеевич
bolshakov34@mail.ru
iabolshakov@gmail.com