СКОЛЬКО СТРАНИЦ НА ДАННОМ ЯЗЫКЕ СОДЕРЖИТ ИНТЕРНЕТ?

И. А. Большаков, С. Н. Галисия-Аро
Центр Компьютерных Исследований
Национальный Политехнический Институт (*IPN*)
Мексика, 07738, г. Мехико
{igor,sofia}@cic.ipn.mx

Ключевые слова: *поисковая машина Гугл, веб-страницы, количество страниц, статистические оценка, метод максимума правдоподобия*.

Утверждается, что для некоторых приложений вычислительной лингвистики необходимо знать полное числа веб-страниц, хранимых в данный момент для выбранного естественного языка в данной поисковой машине Интернета. Показывается, что ряд элементарных шагов по получению статистических сведений о БД поисковой машины Гугл (Google) дают явно противоречивые результаты для простейших теоретико-множественных операций поиска. Игнорируя эти несовершенства, мы предлагаем метод оценки полного числа страниц Гугла для данного языка в данный момент времени. Берутся служебные словоформы, наиболее частотные в некотором представительном текстовом корпусе, переупорядочиваются в соответствии с их представленностью в Гугле, а затем совершаются операции |максимально правдоподобной оценки общего числа страниц исходя из вклада наиболее частотных слов. Метод применен к русской части Гугла и дает результат высокой теоретически предсказываемой точности, которая превышает надежность исходных цифр Гугла.

1. Введение

В подавляющем большинстве случаев Поисковые Машины Интернета (ПМИ) используются ныне для получения информации коммерческого, политического или общепознавательного характера. Но все чаще инженеры ищут с помощью ПМИ сведения о различных изделиях, а ученые – о научных публикациях, авторах и событиях. В то же время использование Интернета для нужд компьютерной лингвистики только началось, см., например, [1, 2, 7], хотя из-за постоянного роста объемов хранимых массивов текстов такое приложение Интернета вполне разумно и актуально. Например, можно себе представить полуавтоматическое пополнение через Интернет словарей конкретного современного языка, автоматической компиляции корпуса текстов [4], выяснение превалирующего начертания и значения того или иного современного термина в рамках интерактивной системы подготовки текста, разрешения неоднозначности данного слова [1], или проверки частоты совместной встречаемости данной пары слов как коллокации [2] – в рамках автоматической обработки текста.

Очевидно, что для компьютерной лингвистики важны не абсолютные цифры встречаемости тех или иных языковых объектов, а их относительные величины. Абсолютное значение встречаемости – это число появления искомого объекта в доступных через Интернет текстах или, на худой конец, число веб-страниц, которые этот объект содержат. Для вычисле-

ния же относительной же величины нужно знать общий объем корпуса текстов, доступного через ту или иную ПМИ, либо же суммарное число веб-страниц, содержащих этот объект.

Настоящая работа ставит целью:

- Пояснить детальнее, почему необходимо знать общий объем текстовых массивов, доступных в конкретный момент для конкретного языка через конкретную ПМИ;
- Рассмотреть одну конкретную ПМИ – Гугл (Google) – в качестве источника сырой статистической информации для определения общего объема тектов выбранного языка;
- Разработать метод оптимальной статистической оценки указанного объема; и
- Применить предложенный метод к массивам, которые Гугл считает русскоязычными.

Мы оцениваем число русскоязычных страниц в Гугле, даем их вычисленную точность и делимся теми сомнениями о достоверности исходных используемых данных, которые возникают при работе с данным ПМИ.

2. Зачем нужно знать суммарный объем данных?

Предположим, что обследуется обширный корпус текстов на предмет пополнения машинного словаря новыми словами. В данной работе не важно, идет ли речь о словаре словоформ или лексем. Тогда в качестве приближенного значения вероятности появления слова W следует брать эмпирическую частоту этого слова, т.е. отношение числа появления $N(W)$ данного слова в обследованном тексте к его длине $N_{\max}$. Слово разумно вставить в словарь, когда эмпирическая частота превышает некий порог P:

$$N(W) / N_{\max} > P. \qquad (1)$$

Ясно, что просуммировав $N(W)$ по всем W мы получим $N_{\max}$, т.е. будет выполнено известное для вероятностей условие нормировки.

Если же нас интересуют устойчивые словосочетания (коллокации), то известным статистическим критерием того, что слова V и W встречаются в текстовом корпусе вместе неслучайно часто, является превышение некоторого порога Q значением так называемой взаимной информации двух этих слов [5]:

$$\ln \frac{N(V,W)}{N_{\max}} > \ln \frac{N(V)}{N_{\max}} + \ln \frac{N(W)}{N_{\max}}, + \ln Q. \qquad (2)$$

Здесь $N(V,W)$ – число случаев совместного (в окне заданной длины) выпадение двух слов, а $N(V)$ и $N(W)$ – числа выпадений каждого из слов, подсчитанные независимо.

Приведенные формулы можно прямо применить к поисковой машине Интернета, дающей статистику слов, которые появляются в рассматриваемых документах в целом. Но чаще ПМИ дает лишь число документов, содержащих исследуемый лингвистический объект хотя бы раз и, быть может, иные объекты того же рода. В итоге отношение $N(W) / N_{\max}$, где N отображают количество документов, а не слов, уже не являются эмпирическими вероятностями и не подчиняются условию нормировки. Однако, между соответствующими вероятностями и относительными долями документов, в которых встречается заданный объект, сохраняется прямая монотонная зависимость, и за неимением лучшего можно использовать формулы (1) и (2) как критерии принятия/непринятия того или иного слова или словосочетания.

Важно то, что в обеих формулах стоит нормирующий числитель N_{max}, который следует понимать как общее число веб-страниц находящихся под управлением данной ПМИ. Без знания этого числа никакие пороговые критерии статистического характера работать не могут, и для задач вычислительной лингвистики эту величину нужно как-то уметь оценивать.

3. Гугл как источник сырой статистики

Приступая к статистическим оцениваниям, основанным на «сырых» данных ПМИ Гугл, мы сразу сталкиваемся к результатами, приводящими в недоумение.

Возьмем две русских словоформы, высокочастотные в больших корпусах текстов, например, *и* и *в*, и сформируем из них несколько запросов к Гуглу, включающих простейшие теоретико-множественные операции. Легко получить следующие количества *N* страниц для различных запросов:

$N(и) = 7\ 150\ 000$; $N(в\ \neg и) = 1\ 830\ 000$; $N(и) + N(в\ \neg и) = 8\ 980\ 000$;

$N(в) = 7\ 110\ 000$; $N(и\ \neg в) = 2\ 070\ 000$; $N(в) + N(и\ \neg в) = 9\ 180\ 000$;

$N(и\ \text{OR}\ в) = 7\ 260\ 000$; $N(и\ \text{OR}\ и) = 7\ 220\ 000$.

Здесь знак отрицания ¬ означает, что берутся страницы, не содержащие последующего слова, а OR означает неисключающую дизъюнкцию появления левой и правой частей. Отсюда вытекает, что:

- Дизъюнктивная операция с точностью до ±0,07% относительно среднего значения не зависит от порядка входящих операндов, что представляется естественным.
- Два математически равнозначных пути конъюнктивных (без применения OR) вычислений количества страниц, которые содержат либо *и*, либо *в*, либо оба эти слова, дают разультаты, различающиеся между собой примерно на 2%, что тоже приемлемо.
- Но дизъюнктивная операция дает результат, отличающийся от среднего для двух конъюнктивных на –20,3% !

Противоречия того же характера, но еще более обескураживающие в части зависимости результата от порядка операндов, встретились при оценке числа испаноязычных страниц в Гугле [3]. Это свидетельствует, что один или оба из указанных путей вычислений содержат серьезную погрешность при теоретико-множественном формировании Гуглом своих данных.

Особые подозрения вызывает дизъюнктивный путь. Действительно, легко получить разительное противоречие

$N(и\ \text{OR}\ в\ \text{OR}\ не) = 7\ 530\ 000$, а $N(и\ \text{OR}\ в\ \text{OR}\ не\ \text{OR}\ с) = 7\ 330\ 000$,

т.е. увеличение числа членов дизъюнктивной формулы ведет к уменьшению суммы! Поэтому мы вообще не используем далее дизъюнкций, а лишь конъюнкции с положительными («содержится») и отрицательными («не содержится») членами-словами.

Этим недостатки Гугла не исчерпываются. Он плохо определяет разные языки. Не имеются в виду случаи, когда заголовок или резюме домумента на веб-странице составлены на одном языке, а остальной текст – на другом. Здесь ошибки вполне естественны. Больше беспокоят ошибочные атрибуции, когда множество целиком болгарских, украинских или белорусских документов признается русскими. Не имея соответствующей статистики и спо-

соба ее добыть, мы этими явлениями вынуждены пренебрегать, считая, что Россия как самая крупная страна, пользующая кириллицей, заметно мажорирует все иные источники с примерно тем же алфавитом.

В любых языках весьма дезориентирующе влияют на статистику страницы, которые по существу лишены текста (объявления, баннеры и прочие странички с картинками, которые могут содержать слова, но нарисованные). К счастью для исследователей русских массивов, страниц такого рода на русском языке пока относительно немного. Хорошо и то, что язык описания структуры отдельных страниц Интернета (а эти части тоже считаются Гуглом текстами) включает латинские операторы и очень редко – литералы в виде русских слов.

Наконец, надо помнить, что все статистические данные Гугла меняются от дня ко дню и даже от часа к часу. Монотонное нарастание полного объема базы данных здесь не причем – результаты зависят от обслуживающего данного пользователя процессора, входящего в поисковую машину, и того пути, которым он обходит массивы данных.

При всех указанных ограничениях, наша цель – преложить метод, пренебрегающий всеми указанными огрехами Гугла и предполагающий, что все его данные абсолютно точны.

4. Метод оптимального оценивания

Предлагаемый метод можно сформулировать следующим образом.

1. Берутся несколько десятков слов данного языка, наиболее частотных в большом текстовом корпусе. Это служебные словоформы – предлоги, союзы, формы вспомогательных глаголов.
2. Для всех указанных форм находятся количества страниц Гугла, на которых эти словоформы встречаются. Формы переупорядочиваются в соответствии с полученной статистикой, и затем оставляется меньшее их число с наивысшими рангами. Для дальнейшего оказались достаточными $K_{max} = 24$ формы.
3. В качестве начального приближения к полному оцениваемому количеству N_{total} страниц берется число $N_1 = N(W(1))$ для слова $W(1)$ первого ранга в Гугле.
4. Далее начинается следующий цикл вычислений:
   - Ищется слово $W(k_2)$ ($k_2 = 2, 3,..., K_{max}$), для которого достигается $N_2 = \max\{N(W(k_2)\neg W(1))\}$, где учитываются страницы, содержащие $W(k_2)$, но не $W(1)$. Число N_2 прибавляется к N_{total}.
   - Ищется слово $W(k_3)$ ($k_3 = 2, 3,..., K_{max}, k_3 \neq k_2$), для которого достигается $N_3 = \max\{N(W(k_3)\neg W(1)\neg W(k_2))\}$, где учитываются только страницы без $W(1)$ и $W(k_2)$. Число N_2 прибавляется к N_{total} и т. д.

Вклад в общее число страниц, сносимый очередными словами без всех уже учтенных предыдущих мог бы быть подсчитан указанным образом вплоть до $W(K_{max})$ и далее, но Гугл не позволяет использовать более $K = 10$ элементов в запросной формуле. Поэтому остановим вычисления N_{total} на инкременте $\max\{N(W(k_{10}) \neg W(1) \neg W(k_2) \neg W(k_3) ...\neg W(k_9))\}$, т. е. на десятом максимально влиятельном слове, пронумерованном согласно приведенной выше процедуре среди всех взятых K_{max} слов.

В принципе, слова тоже могут быть приняты во внимание, но в предположении, что отбрасывается не более девяти слов, учтенных ранее. Однако такие вычисления дадут заведомо завышенный результат и мы их отвергаем.

Чтобы получить более реалистическую оценку всего прочего шлейфа слов и в то же время оценить стандартное отклонение этой оценки, аппроксимируем последовательность случайных величин N_1, N_2 ..., N_{10} бесконечным экспоненциальным рядом

$$\tilde{N}_{\text{total}} \approx \mathrm{e}^{a} (1 + \mathrm{e}^{-u} + \mathrm{e}^{-2u} + ...) = \mathrm{e}^{a} / (1 - \mathrm{e}^{-u}),$$

где $a >> 1$ и $u > 0$ определяются через имеющиеся N_1, N_2, ..., N_{10} методом максимального правдоподобия [5]. Это дает следующие оценки для a и u:

$$\tilde{a} = \frac{2\mathrm{M}_0(2\mathrm{K}+1)}{\mathrm{K}(\mathrm{K}-1)} - \frac{6\mathrm{M}_1}{\mathrm{K}(\mathrm{K}-1)}, \qquad \tilde{u} = \frac{6\mathrm{M}_0}{\mathrm{K}(\mathrm{K}-1)} - \frac{12\mathrm{M}_1}{\mathrm{K}(\mathrm{K}^2-1)},$$

где $\mathrm{M}_0 = \Sigma_{i=1...\mathrm{K}} \ln(N_i)$, $\mathrm{M}_1 = \Sigma_{i=1...\mathrm{K}} \ln(N_i)\, i$. Для оценки дисперсии σ^2 случайных величин $\ln(N_i)$ справедлива формула

$$\sigma^2 = \frac{1}{\mathrm{K}} \sum_{i=1....\mathrm{K}} (\ln(N_i) - \tilde{a} - \tilde{u}\, i)^2.$$

Чтобы найти дисперсию $\mathbf{D}(\tilde{N}_{\text{total}})$, намеренно игнорируем случайную природу величины σ^2 и вычислим минимальную ковариационную матрицу пары случайных величин $\{\tilde{a}, \tilde{u}\}$ как если бы оценивались только они. Действительно, влияние случайности σ^2 измеряется более высокими степенями этой величины, а дисперсия $\mathbf{D}(\tilde{N}_{\text{total}})$ и так ожидается достаточно малой. В этом приближении сформируем линейную комбинацию элементов ковариационной матрицы и вторых частных производных от $\tilde{N}_{\text{total}}$ по a и u:

$$\mathbf{D}(\tilde{N}_{\text{total}}) = \frac{\partial^2 \tilde{N}_{\text{total}}}{\partial a^2}\, \mathbf{D}(\tilde{a}) + 2\frac{\partial^2 \tilde{N}_{\text{total}}}{\partial a\, \partial u}\, \mathbf{R}(\tilde{a}, \tilde{u}) + \frac{\partial^2 \tilde{N}_{\text{total}} 1}{\partial u^2}\, \mathbf{D}(\tilde{u}),$$

где производные берутся в точке $\{a = \tilde{a}, u = \tilde{u}\}$, а

$$\mathbf{D}(\tilde{a}) = \frac{2\sigma^2(2\mathrm{K}+1)}{\mathrm{K}(\mathrm{K}-1)}\ ;\ \mathbf{R}(\tilde{a}, \tilde{u}) = -\frac{6\sigma^2}{\mathrm{K}(\mathrm{K}-1)}\ ;\ \mathbf{D}(\tilde{u}) = \frac{12\sigma^2}{\mathrm{K}(\mathrm{K}^2-1)}.$$

Окончательно получаем

$$\mathbf{D}(\tilde{N}_{\text{total}}) = \frac{2\sigma^2 e^{a}}{\mathrm{K}(\mathrm{K}-1)(1-e^{-u})}\left[(2\mathrm{K}+1) + \frac{3e^{-u}}{(1-e^{-u})} + \frac{6e^{-u}(1+e^{-u})}{(\mathrm{K}+1)(1-e^{-u})^2}\right],$$

где величины a и u снова берутся оценочными: $\tilde{a}$ и $\tilde{u}$.

После всех вычислений стандартную девиацию $\sqrt{\mathbf{D}(\tilde{N}_{\text{total}})}$ следует сравнить с оцененной величиной $\tilde{N}_{\text{total}}$.

5. Оценка числа русских страниц в Гугле

Для реализации предложенного метода применительно к произвольному естественному языку была построена программа на языке программирования PERL, содержащая обращения к Гуглу. Для каждого их них автоматически образуется запрос в виде формулы, включающей от одного до десяти слов выбранного языка. Нужная цифра числа страниц содержится в одной из начальных строк полученной справочной страницы. В [3] даны результаты применения развитого метода к испанскому языку.

Применительно к русскому языку некоторое затруднение вызвало то, что тексты на кириллице в запросе нужно представлять в юникоде. В варианте UTF-8 этой кодировки каждая русская буква задается тремя байтами – знаком процента и двумя шестнадцатиричными буквами. Запросы формировались из таких троек.

Исходный список составили 75 служебных словоформ из 230 форм, наиболее частотных в корпусе длиной более 3 млн. слов. Для всех их были найдены количества включающих страниц в Гугле, а затем были отобраны 24 наиболее представительных в данной ПМИ. Дальнейшие вычисления велись согласно приведенному выше методу.

В таблице 1 представлены результаты последовательного переупорядочения выбранных слов по данным Гугла. Первый столбец слева задает русскую словоформу, второй дает число содержащих ее миллионов веб-страниц, третий указывает ранг слова в корпусе, а четвертый – его ранг, определенный в процессе вычисления по развитому методу (содержимое последнего столбца будет пояснено ниже).

Таблица 1. Ранги выбранных слов и их представленность в Гугле

Слово	**Число страниц, в млн.**	**Ранг в корпусе**	**Ранг при вычисл.**	**Ранг при вычисл. (2-й экс.)**
и	7.09	**1**	**1**	**2**
в	7.05	**2**	**2**	**1**
не	6.90	**3**	**9**	**9**
с	6.42	**7**	**4**	**4**
на	6.34	**4**	**3**	**3**
уже	6.33	**30**	**–**	**–**
может	6.23	**45**	**–**	**–**
что	6.14	**5**	**–**	**–**
а	6.08	**8**	**7**	**7**
ее	5.98	**37**	**–**	**–**
чтобы	5.84	**48**	**–**	**–**
по	5.77	**16**	**5**	**5**
мы	5.71	**28**	**–**	**–**
они	5.68	**29**	**–**	**–**
когда	5.67	**34**	**–**	**–**
к	5.60	**11**	**6**	**6**
нет	5.60	**42**	**10**	**10**
как	5.58	**10**	**–**	**–**
был	5.58	**35**	**–**	**–**
их	5.58	**36**	**–**	**–**
еще	5.52	**23**	**–**	**–**
этом	5.47	**64**	**–**	**–**
все	5.45	**14**	**8**	**8**
без	5.44	**55**	**–**	**–**

Цикл запросов, составленных из десяти наиболее влиятельных (для наших вычислений) слов, отражен в таблице 2.

Таблица 2. Цикл запросов к Гуглу

Число добавл.	**Запрос с Гуглу**

страниц	
7 090 000	*и*
1 890 000	*в¬и*
822 000	*на¬и¬в*
291 000	*с¬и¬в¬на*
227 000	*по¬и¬в¬на¬с*
116 000	*к¬и¬в¬на¬с¬по*
86 900	*а¬и¬в¬на¬с¬по¬к*
80 900	*все¬и¬в¬на¬с¬по¬к¬а*
63 500	*не¬и¬в¬на¬с¬по¬к¬а¬все*
29 200	*нет¬и¬в¬на¬с¬по¬к¬а¬все¬не*
10 696 500	$= N_{\text{total}}$ **(сумма страниц по десяти словам)**

На рис. 1 даны точки, представляющие N_i в логарифмическом масштабе, а прямая линия изображает их линейную аппроксимацию, полученную методом максимума правдоподобия. Оценка имеет вид $\ln(N_i) = \tilde{a} - \tilde{u}\, i$, где $\tilde{a} = 15{,}41$ и $\tilde{u} = 0{,}54$. Отсюда $\tilde{N}_{\text{total}} =$ 11 840 000. Это означает, что экспоненциальная аппроксимация добавила $\tilde{N}_{\text{total}} - N_{\text{total}} = 1{,}14$ млн. страниц к количеству, накопленному за счет первых десяти слов.

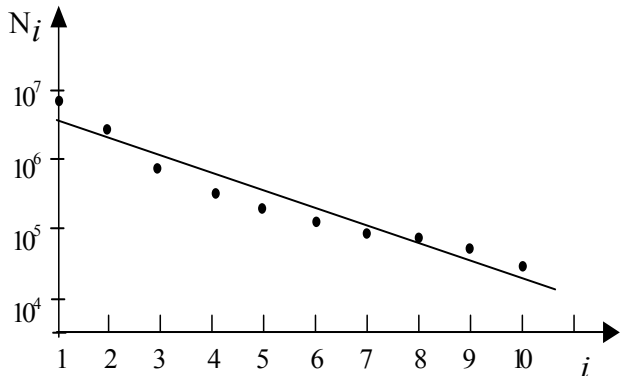

Рис. 1. Аппроксимация вклада отдельных слов

Стандартное отклонение $\sqrt{\mathbf{D}(\tilde{N}_{\text{total}})}$ получилось равным 26 000, что составляет лишь 1/400 от измеренной величины. Оно пренебрежимо мало в сравнении с подозреваемой неточностью исходных статистических данных. Наш инструмент очевидным образом точнее, чем выбранное ему применение.

Попробуем оценить теперь чувствительность нашего метода к тому, слово какого ранга в Гугле взято для начальной аппроксимации. Проведем те же вычисления, но в качестве первого в цикле возьмем слово со вторым рангом. Ранговые данные приведены в последнем столбце таблицы 1, а количественно в этом случае получаем $\tilde{N}_{\text{total}} =$ 12 090 000 при той же дисперсии. Как видим, порядок привлечения страниц не изменился, а различие между двумя окончательными результатами составило примерно 2%, что свидетельствует о достаточной устойчивости нашего метода.

Итак, если верить исходным данным Гугла, на момент оценки он содержал примерно 11,9 млн. русских страниц. В то же время полный объем БД Гугла превышает 3 млрд. страниц, и примерно на 90% это тексты на английском языке.

6. Заключение

Предложен метод оценки максимального правдоподобия числа страниц с тестами на конкретном естественном языке в поисковых машинах Интернета. Его применение к одной из самых мощных по объему хранимой информации машин – Гуглу – показал, что на момент оценки здесь хранилось примерно 11,9 млн. русских страниц.

Наш метод легко переносим и на другие поисковые машины с русскими текстами. Для этого достаточно, чтобы они выдавали по запросу общее число вхождений словоформ самых высоких рангов и/или общее количество страниц, включающих данное слово, а также выполняли простейшие теоретико-множественные операции по приведенной выше схеме. Наша оговорка об оценивании нужных слов принципиальна, поскольку идет речь о чисто вспомогательных словах, неинтересных для большинства иных типов поисковых операций.

Благодарности

Настоящая работа выполнена при частичной поддержке мексиканских правительственных организаций CONACyT, CGEPI-IPN и SNI. Выражаем признательность проф. А.Ф. Гельбуху за советы и проф. Г. О. Сидорову за предоставление статистических данных по корпусу русских текстов.

Литература

1. Agirre, E., D. Martinez. *Exploring automatic word sense disambiguation with decision lists and the Web*. Proceedings of Semantic Annotation and Intelligent Annotation Workshop. Organized by COLING, Luxemburg, 2000.

2. Bolshakov, I. A. *Detección y corrección de malapropismos en español mediante un sistema bietapa para comprobar colocaciones*. Memoria del Congreso Internacional de Computación "Avances en Ciencias de Computación e Ingeniería de Cómputo" CIC'2002. CIC-IPN, Mexico, Noviembre, 2002, p.304-313.

3. Bolshakov, I. A., Sofía N. Galicia-Haro. *Can We Correctly Estimate the Total Number of Pages in Google for a Specific Language?* In: A. Gelbukh (Ed.). *Computational Linguistics and Intelligent Text Processing*. Proc. Intern. Conf. on Computational Linguistics CICLing-2003, February 2003, Mexico City. Lecture Notes in Computer Science No. 2588, Springer, 2003, p. 415-419.

4. Gelbukh, A., G. Sidorov, and Liliana Chanona-Hernández. *Compilation of a Spanish representative corpus.* In: A. Gelbukh (Ed.). *Computational Linguistics and Intelligent Text Processing*. Proc. Intern. Conf. on Computational Linguistics CICLing-2002, February 2002, Mexico City. Lecture Notes in Computer Science No. 2276, Springer, 2002, p. 285–288.

5. Крамер, Г. *Математические методы статистики.* М.: «Мир». 1975.

6. Keller, F., M. Lapata, and O. Ourioupina. *Using the Web to Overcome Data Sparseness*. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Philadelphia, Pennsylvania, USA. 2002.

7. Manning, Ch. D., H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

How Many Pages in this Language does Internet contain?

I. A. Bolshakov, S. N. Galicia-Haro

Keywords: *search engine Google, web pages, page amount, statistical estimate, maximum likelihood method*.

It is argued that for some applications of computational linguistics the total amount of web-pages actually stored in an Internet search engine for a specific language is relevant. It is shown that some elementary steps in getting statistics characterizing Google engine's DB are bewildering: simple set theory operations give evidently inconsistent results. Ignoring these imperfections, we propose a method for estimation of the total page amount for a given language. It takes functional words most frequent in a representative text corpus for a given language, reorders them according to their availability in Google, and then compute the maximum likelihood estimate of the total amount through contributions of the high-rank words. The method is applied to Russian pages and gives results with a high precision exceeding those of the source statistical data.