«Информативность частотных характеристик *N*-грамм текстовых фрагментов интернет-сайтов для поисковых систем»

В.А. Строцев

Поиск информации в Интернет-среде уже невозможно представить без использования поисковых систем. В настоящее время в них реализуются разнообразные алгоритмы и принципы поиска, при этом процесс совершенствования таких систем реализуется уже с 1994 года (с момента открытия первого проекта каталога сайтов для организации доступа к информационным ресурсам сети – сайт Yahoo.com). Тем не менее, пользователи не всегда удовлетворены результатами обращения к ним. Ряд интересных фактов, связанных с взаимоотношением пользователей и поисковых систем приведены на сайте [1]: пользователи бросают поиски после 12 минут бесплодных попыток; около 75% пользователей разочаровываются при поиске информации в Интернете.

Отметим, что качество предоставляемых пользователю ответов в существенной мере зависит от сформированного запроса. Однако в силу ряда обстоятельств пользователь не всегда в состоянии достаточно точно сформулировать запрос и количество полученных им ответов становится большим. В этих ситуациях дополнительным признаком отбора релевантных ответов может являться принадлежность текстовых документов к той или иной неявной группе. Неявность группы проявляется в том, что принадлежность текста к ней определяется не прямым сравнением с эталонными (ключевыми) словами, а по соответствию смысловым признакам, формулировка которых в искомом тексте отсутствует. Например, пользователь хочет найти описание сказочного персонажа – летающей собаки с именем «Фалькорн». Он точно знает, что это персонаж художественного произведения, автора и название которого он не помнит. По ключевым словам «Фалькорн», «летающая собака» число ссылок очень велико и их просмотр утомителен (возможно займёт более 12 минут, что неминуемо классифицирует эту попытку обращения к поисковой системе как неуспешную). При добавлении ключевых слов «художественное произведение» или «сказка» ситуация существенно не улучшается, поскольку в самом произведении (сайте, содержащем искомую информацию) этих слов может и не быть. Дополнительный отслеживаемый признак позволил бы значительно сократить число сайтов-результатов поиска, отсекая информацию просто о летающих собаках.

Такая оценка принадлежности текстовых фрагментов интернет-сайтов к выбранной неявной группе может быть реализована на основе аппарата теории математической лингвистики [2], которая изучает закономерности лингвистических объектов.

Относительно рассматриваемого направления следует выделить работы, в которых для решения практических задач применяется устойчивость частот отдельных символов и их сочетаний заданной длины *N* (*N*-грамм). Так в работе В. Канвара и Дж. Тренкла [3] был предложен метод определения языка документа, основанный на сравнении частот *N*-грамм текста с их частотами для различных языков. В работе [4] *N*-граммы уровня символов применены для семантической классификации незнакомых собственных имён, а в статье [5] анализируется содержание и применение *N*-грамм как средства фиксации языковых реалий и показывается соотношение моделей *N*-грамм, формальной грамматики и теории случайных марковских процессов. Делается вывод о широких возможностях таких моделей для автоматического анализа печатных текстов. Следует отметить, что в теории поиска как физических, так и информационных объектов также широко применяются марковские модели [6] – [10]. Но до их применения следует сначала оценить информативность соответствующих признаков.

Однако информативность частотных характеристик *N*-грамм текстовых фрагментов интернет-сайтов для формирования дополнительного признака их принадлежности к неявным группам для совершенствования поисковых систем ещё не рассматривалась.

Целью работы является оценка возможности применения частотных характеристик *N*-грамм текстовых фрагментов интернет-сайтов для совершенствования поисковых систем на основе исследования их информативности.

Постановка задачи

Пусть определён корпус текстовых документов (фрагментов текстов, являющихся содержанием страниц интернет-сайтов) общим объёмом *M* знаков, распределённых по *V* темам (неявным группам). Суммарные объёмы фрагментов, относящихся к каждой *n*-ой теме ($n = \overline{1,V}$) известны, и соответственно равны m_n, $\sum_{n=1}^{N} m_n = M$. Значения m_n, $n = \overline{1,V}$ являются значительными для оценки частотных характеристик каждого общего текста темы (неявной группы).

Требуется оценить информативность частотных характеристик текстовых фрагментов интернет-сайтов для формирования дополнительного признака их принадлежности неявным группам для совершенствования поисковых систем.

Разработка методики исследования

На подготовительном этапе для выбранного языка составляются возможные последовательности значащих символов системы письменности длиной не более *N*: ξ_i^N, $i = \overline{1, I(N)}$. К незначащим символам можно отнести цифры, знаки пунктуации, пробелы и т.п. Если число значащих символов системы письменности обозначить через Ω, то число возможных последовательностей значащих символов длиной не более *N* для этой системы письменности определяется по выражению:

$$I(N) = \sum_{n=1}^{N} \Omega^n . \qquad (1)$$

Функциональная зависимость в выражении (1) представлена только от *N*, поскольку реального механизма влияния на Ω не имеется.

Для каждой *n*-ой неявной группы подсчитывается число использований *N*-грамм ξ_i^N: $\nu_n(\xi_i^N)$.

Тогда групповые частоты определяются по выражениям вида:

$$p_n(\xi_i^N) = \frac{\nu_n(\xi_i^N)}{\sum_{i=1}^{I(N)} \nu_n(\xi_i^N)}, \; n = \overline{1,V}, \; i = \overline{1, I(N)}. \qquad (2)$$

Применение выражения (2) подразумевает знание (подсчёт) числа использований всех $I(N)$ последовательностей. Однако поскольку в приложениях используется ограниченное число наиболее употребительных *N*-грамм $\bar{I} \le I(N)$ (для идентификации языка текста в соответствии с [3] – не более 300), то вычисления по выражению (2) требуют использование неоправданно больших ресурсов (для $\Omega = 26$ величина $I(N)$ для $N = 3,4,5$ принимает значение $I(3) = 18\,278$, $I(4) = 475\,254$, $I(5) = 12\,356\,630$).

Более «экономным» с вычислительной точки зрения является применение относительных частот вида

$$p_n^*(\xi_i^N) = \frac{\nu_n(\xi_i^N)}{m_n}, \; n = \overline{1,V}, \; i = \overline{1, I(N)}. \qquad (3)$$

Проверка гипотезы о возможности такой замены приведена в экспериментальной части статьи.

Более того, поскольку наиболее употребительные *N*-граммы в каждой неявной группе могут породить различные наборы последовательностей, а для реализации

сравнительных процедур, как правило, требуется использование соотносимых наборов, то для определения отсортированного по убыванию набора $\bar{I}$ N-грамм ϕ_j^N, $j=\overline{1,\bar{I}}$ для заданного корпуса текстовых документов требуется выполнение процедуры следующего вида:

$$\phi_1^N = \arg\max_{i=1,I(N)}\left(\sum_{n=1}^{V}\nu_n(\xi_i^N)\right), \qquad \phi_2^N = \arg\max_{\substack{i=1,I(N),\\ \xi_i^N\neq\phi_1^N}}\left(\sum_{n=1}^{V}\nu_n(\xi_i^N)\right),$$

$$\phi_3^N = \arg\max_{\substack{i=1,I(N),\\ \xi_i^N\neq\phi_1^N,\\ \xi_i^N\neq\phi_2^N}}\left(\sum_{n=1}^{V}\nu_n(\xi_i^N)\right),\ldots,\ \phi_{\bar{I}}^N = \arg\max_{\substack{i=1,I(N),\\ \xi_i^N\neq\phi_1^N,\\ \xi_i^N\neq\phi_2^N,\\ \ldots\\ \xi_i^N\neq\phi_{\bar{I}-1}^N}}\left(\sum_{n=1}^{V}\nu_n(\xi_i^N)\right). \qquad (4)$$

Тогда с учётом (4) относительные частоты соотносимых наборов N-грамм могут быть получены по выражениям:

$$p_n^*(\phi_j^N)=\frac{\nu_n(\phi_j^N)}{m_n},\ n=\overline{1,V},\ j=\overline{1,\bar{I}}. \qquad (5)$$

Пусть некоторый l-ый текстовый фрагмент интернет-сайта требуется отнести к одной из V неявных групп. Объём этого фрагмента составляет m_l^t знаков. Тогда относительные частоты соотносимых наборов N-грамм для этого фрагмента вычисляются по выражениям:

$$q_l^*(\phi_j^N)=\frac{\nu_l^t(\phi_j^N)}{m_l^t},\ j=\overline{1,\bar{I}}, \qquad (6)$$

где $\nu_l^t(\phi_j^N)$ – число использований N-граммы ϕ_j^N в l-ом текстовом фрагменте.

По полученным значениям относительных частот (5) и (6) можно организовать процедуру сравнения и оценки принадлежности l-го текстового фрагмента интернет-сайта к одной из V неявных групп.

Одним из наиболее простых способов её организации является:

1. Расчёт выборочных коэффициентов корреляции Пирсона [11], с. 128:

$$r_{ln}^{tN}(p_n^*,q_l^*))=\frac{\bar{I}\cdot\sum_{j=1}^{\bar{I}}p_n^*(\phi_j^N)\cdot q_l^*(\phi_j^N)-\sum_{j=1}^{\bar{I}}p_n^*(\phi_j^N)\cdot\sum_{j=1}^{\bar{I}}q_l^*(\phi_j^N)}{\sqrt{\left[\bar{I}\cdot\sum_{j=1}^{\bar{I}}\left(p_n^*(\phi_j^N)\right)^2-\left(\sum_{j=1}^{\bar{I}}p_n^*(\phi_j^N)\right)^2\right]\cdot\left[\bar{I}\cdot\sum_{j=1}^{\bar{I}}\left(q_l^*(\phi_j^N)\right)^2-\left(\sum_{j=1}^{\bar{I}}q_l^*(\phi_j^N)\right)^2\right]}}, \qquad (7)$$

$n=\overline{1,V}$.

2. Принятие решения о принадлежности текстового фрагмента Φ_l к одной из неявных групп T_n, $n=\overline{1,V}$ в соответствии с правилом:

$$\Phi_l\in T_{n^*},\ n^*=\arg\max_n r_{ln}^{tN}(p_n^*,q_l^*)). \qquad (8)$$

Естественно, что могут решаться и другие задачи, например, проверка статистической гипотезы о значимости коэффициентов корреляции, равенстве их между собой и т.д.

Результаты исследования

Проведём экспериментальное исследование в соответствии с разработанной методикой. В качестве источника фрагментов текстов, являющихся содержанием страниц интернет-сайтов, выберем англоязычный сайт [12], на котором представлены материалы

по различным темам. Выберем четыре темы: «Computers & Internet», «Music and Movies», «Pets and Animals» и «Politics and Government» и поставим им в соответствие значение n в порядке перечисления. Примем значение N, равное 3 (в работе [3] N принимает значения от 1 до 5).

Для сформированного корпуса: $\Omega = 26$, $V = 4$, $m_1 = 11\,192\,104$, $m_2 = 7\,737\,926$, $m_3 = 10\,862\,615$, $m_4 = 3\,767\,664$, $M = 33\,560\,309$, $\sum_{i=1}^{I(3)} \nu_1(\xi_i^3) = 21\,071\,631$, $\sum_{i=1}^{I(3)} \nu_2(\xi_i^3) = 14\,041\,628$, $\sum_{i=1}^{I(3)} \nu_3(\xi_i^3) = 19\,980\,994$, $\sum_{i=1}^{I(3)} \nu_4(\xi_i^3) = 7\,077\,912$.

Для оценки качества последовательностей были рассчитаны выборочные парные коэффициенты корреляции между различными парами множеств $\left\{\nu_n(\xi_i^N), i = \overline{1,\bar{I}}\right\}$, $n = \overline{1,V}$, $r_{n_1 n_2}(\bar{I})$, $n_1 = \overline{1,V-1}$, $n_2 = \overline{n_1+1,V}$, представленные в таблице № 1.

Таблица № 1

Выборочные парные коэффициенты корреляции

$\bar{I}$	$r_{12}(\bar{I})$	$r_{13}(\bar{I})$	$r_{14}(\bar{I})$	$r_{23}(\bar{I})$	$r_{24}(\bar{I})$	$r_{34}(\bar{I})$
18278	0,997207	0,996013	0,996548	0,997745	0,997745	0,995870
1000	0,997075	0,996412	0,995821	0,997653	0,997667	0,995686
500	0,997169	0,995801	0,996403	0,997733	0,997743	0,995610
400	0,997216	0,995784	0,996374	0,997793	0,997772	0,995560
300	0,997221	0,995759	0,996347	0,997802	0,997769	0,995478
200	0,997191	0,995767	0,996212	0,997962	0,997769	0,995470
100	0,997039	0,995454	0,995897	0,997966	0,997692	0,995265

Максимальное относительное отклонение

$$\delta r^{\max} = \max_{\substack{n_1=\overline{1,V-1},\\ n_2=\overline{n_1+1,V}}} \left(\frac{\max\limits_{\bar{I}\in\tilde{I}} r_{n_1 n_2}(\bar{I}) - \min\limits_{\bar{I}\in\tilde{I}} r_{n_1 n_2}(\bar{I})}{\max\limits_{\bar{I}\in\tilde{I}} r_{n_1 n_2}(\bar{I})} \right) \cdot 100\%,$$

$\tilde{I} = \{18278, 1000, 500, 400, 300, 200, 100\}$, при уменьшении $\bar{I}$ с $\bar{I} = I(3) = 18\,278$ до $\bar{I} = 100$ составило 0,096%.

Таким образом, косвенно подтверждается гипотеза о возможности существенного ограничения числа рассматриваемых наиболее употребительных N-грамм $\bar{I}$ при решении прикладных задач.

Для оценки принадлежности произвольного l-го текстового фрагмента Φ_l на тему «Computers & Internet» к одной из неявных групп T_n, $n = \overline{1,4}$ и исследования информативности частотных характеристик N-грамм положим, что

$$q_l^*(\phi_j^N) = p_1^*(\phi_j^N) + \overline{\mathrm{N}}_j[0, \varepsilon \cdot p_1^*(\phi_j^N)],\ j = \overline{1,\bar{I}}, \qquad (9)$$

где $\overline{\mathrm{N}}_j[0, \varepsilon \cdot p_n^*(\phi_j^N)]$ – обозначение j-ой случайной величины, распределённой по нормальному закону с нулевым математическим ожиданием и средним квадратичным отклонением (СКО) $\varepsilon \cdot p_n^*(\phi_j^N)$, ε – параметр вариации, $\varepsilon \in \left[0, \frac{p_n^*(\phi_j^N)}{3}\right]$.

Отметим, что при моделировании частотных характеристик N-грамм текстовых фрагментов относительно выражения (2) в соответствии с подходом, определяемым выражением (9), значения выборочных коэффициентов корреляции, рассчитанные по выражению, соответствующему (7), оказались равны аналогичным выборочным коэффициентам упрощённой модели (выражение (3)).

Результаты оценки вероятности неправильного решения о принадлежности текстовых фрагментов с частотными характеристиками *N*-грамм, полученных по выражению (9), для различных значений ε и $\bar{I}$ при числе реализаций моделирования случайных величин равном 100 (100 различных фрагментов) и процедуре принятия решения (7), (8), представлены в таблице №2.

Таблица №2

Оценки вероятностей неправильного решения о принадлежности текстовых фрагментов

$\bar{I}$	$\varepsilon \le 0{,}05$	$\varepsilon = 0{,}10$	$\varepsilon = 0{,}15$
18278	0,00	0,01	0,05
1000	0,00	0,01	0,06
500	0,00	0,02	0,14
400	0,00	0,01	0,11
300	0,00	0,01	0,14
200	0,00	0,02	0,16
100	0,00	0,03	0,21

Из анализа таблицы видно, что текстовые фрагменты надёжно классифицируются при величинах СКО составляющих практически до 10% от значений относительных частот соответствующих *N*-грамм. При этом уменьшение числа рассматриваемых отсортированных по убыванию относительной частоты *N*-грамм существенно сказывается только для величин СКО превышающих 10% от значений относительных частот этих *N*-грамм.

Заключение

Использована закономерность математической лингвистики: каждый из символов встречается в тексте с определенной частотой и обладает особыми валентностями, т. е. лингвистическими способностями сочетаться с другими символами [2]. Отметим, что рассматриваемая методика обладает большой общностью в отношении систем письменности, поскольку не опирается только на алфавитные системы.

Выводы:

1. Частотные характеристики *N*-грамм текстовых фрагментов интернет-сайтов обладают достаточной степенью информативности для совершенствования поисковых систем на их основе.

2. Существует неравномерное распределение зависимости информативности частотных характеристик *N*-грамм текстовых фрагментов интернет-сайтов от неявных групп (в условиях рассмотренного примера более различимыми оказались пары тем «Computers & Internet»–«Pets and Animals», «Computers & Internet»–«Politics and Government» и «Pets and Animals»–«Politics and Government», т.е. важной задачей является выбор и описание соответствующей неявной группы.

Литература

1. Я мыслю, следовательно, раскручиваю // Исследования и статистика в области интернета, интернет рекламы и продвижения сайта. [Электронный ресурс]: http://digits.ru (дата обращения: 20.12.2012).

2. Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. Учеб. пособ. М.: Высш. шк. 1977. – 383 с.

3. Cavnar W. B., Trenkle J. M. N-Gram-Based Text Categorization // In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications. 1994.

4. Нехай И.В. Применение N-грамм и других статистик уровня символов и слов для семантической классификации незнакомых собственных имён // Международная

конференция по компьютерной лингвистике. [Электронный ресурс]: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/150.pdf (дата обращения: 20.12.2012).

5. Гудков В.Ю., Гудкова Е.Ф. N-граммы в лингвистике // Вестник Челябинского университета. – 2011. – №24 (239). Филология. Искусствоведение. – Вып. 57. – С. 69 – 71.

6. Строцев А.А. Иващенко И.Л. Синтез оптимального управления многопозиционной информационной системой при поиске группы динамических объектов // Известия высших учебных заведений. Радиоэлектроника. – 2005. – Т.48. – №10. – С. 37–45.

7. Строцев А.А. Совместное оптимальное управление поиском и наблюдениями за условно детерминированными динамическими объектами в импульсной многоканальной измерительно-поисковой системе // Известия высших учебных заведений. Радиоэлектроника. – 2004. – Т.47. – №9. – С. 22–29.

8. Строцев А.А. Оптимизация поиска и наблюдений многоканальной импульсной радарной станции в составе многопозиционной комплексной измерительно-поисковой системы // Автоматика и вычислительная техника. – 2004. – №3. – С. 12–21.

9. Развитие PageRank // [Электронный ресурс]: http://ornitos.blogspot.ru (дата обращения: 20.12.2012).

10. Грищук Т.В. Получение характеристической обсервации скрытой марковской модели // Наукові праці ВНТУ. – 2007. – № 1.

11. Третьяк Л.Н. Обработка результатов наблюдений. Оренбург: ГОУ ОГУ, 2004. – 171 с.

12. ArticleCity.com // Free Articles For Reprint. [Электронный ресурс]: http://www.articlecity.com (дата обращения: 20.12.2012).