И. З. Батыршин, Р. Н. Гильмутдинов, О задаче поиска информации в Интернет, *Исслед. по информ.*, 2004, выпуск 8, 129–138

О ЗАДАЧЕ ПОИСКА ИНФОРМАЦИИ В ИНТЕРНЕТ

И.З. Батыршин, Р.Н. Гильмутдинов

Возможность эффективного поиска и извлечения информации имеет важнейшее значение для использования всего потенциала Интернет. Процесс развития сети Интернет и увеличение объема содержащихся в ней данных делает задачу информационного поиска все более актуальной.

По оценкам некоммерческой общественной корпорации «Internet Systems Consortium, Inc.» (ISC), в январе 2004 года в Интернет было зарегистрировано более 233 тысяч доменных имен [1]; трафик (количество передаваемой информации) увеличивается каждый месяц на 30%. По прогнозу Computer Industry Almanac Inc. к 2005 году число пользователей увеличится до 945 миллионов [2].

Наблюдаемое лавинообразное разрастание Интернет, которое, по всей видимости, еще будет продолжаться ближайшие 10-20 лет, привело к тому, что существующие информационно-поисковые сервисы и системы (ИПС) фактически не справляются с задачей поиска информации. Так, по некоторым исследованиям, ни одна из глобальных ИПС не покрывает более чем 16% из содержащихся во всемирной сети страниц. Кроме того, достаточно низкой остается эффективность поиска.

Связано это в первую очередь с организацией процесса поиска в ИПС [6, 10, 11, 13-16]. Поиск информации в Интернет – частный случай задачи информационного поиска. Такая задача определяется совокупностью объектов поиска – источников данных (например, книги, журналы и газеты в библиотеке; таблицы в базе данных), и совокупностью субъектов (например, посетитель библиотеки), осуществляющих поиск необходимых данных. В контексте задачи информационного поиска в Интернет под объектом будет пониматься отдельный документ (веб-страница), характеризующийся уникальным Интернет-адресом (URL); под субъектом – пользователь, выполняющий запрос в целях поиска конкретного подмножества документов.

Модель задачи поиска информации

Рассмотрим принципы работы гипотетической информационно-поисковой системы и на основе анализа принципов формализуем задачу поиска информации в Интернет.

Любая ИПС состоит из следующих элементов [17]: информационно-поисковый язык (ИПЯ); правила перевода текстов документов и запросов с естественного языка на ИПЯ; алгоритмы поиска; технические устройства, реализующие алгоритмы поиска; база документов (или их адресов), размещенных на каких-либо носителях информации.

Обозначим все множество размещенных в сети документов через $\overline{UD} = \left\{\overline{Ud}^i\right\}_{i=1}^{\overline{N}}$, где $\overline{N}$ – количество документов.

Поиск информации основан на индексировании размещенных в Интернет документов, то есть процессе создания представления документа путем ассоциирования с ним дескрипторов содержимого (терминов). Впоследствии термины используются для оценки релевантности документа запросу пользователя, что имеет непосредственное отношение к эффективности извлечения информации ИПС.

Под релевантностью документа понимается соответствие смыслового содержания документа запросу пользователя [17].

Для рассматриваемой ИПС определим множество проиндексированных документов как $UD = \left\{Ud^i\right\}_{i=1}^{N}$, где $UD \subset \overline{UD}$; $N = \alpha * \overline{N}$, $0 < \alpha \le 1$.

Различают два типа терминов: объективные и необъективные. Объективные термины – это термины, присущие семантическому содержимому документа, выбор объективных дескрипторов содержимого тривиален. К подобным терминам относятся фамилия автора, Интернет-адрес документа, дата его публикации. Необъективные термины отражают информацию самого документа, также их называют «терминами содержимого». Однозначные правила выбора необъективных терминов, а также правила определения степени их соответствия пока не предложены.

Определим множества объективных и необъективных терминов как $TOb = \left\{Tob^i\right\}_{i=1}^{Nob}$ и $TSub = \left\{Tsub^i\right\}_{i=1}^{Nsub}$ соответственно, где Nob и $Nsub$ – размерности множеств.

Обычно при индексировании с документами соотносятся необъективные термины. Такое ассоциирование может дополняться назначением весового коэффициента, определяющего степень представления или отражения данным термином содержимого документа.

Для рассматриваемой ИПС определим оператор индексирования Ind, который каждому документу Ud^i ставит в соответствие вектор терминов $T^i = \left(Tob^i, Tsub^i\right) = \left(Tob^i_1, \ldots, Tob^i_{M^i ob}, Tsub^i_{M^i ob+1}, \ldots, Tsub^i_{M^i ob+M^i sub}\right) = \left(T^i_1, \ldots, T^i_{M^i}\right)$, где $Tob^i_j \in TOb$, $Tsub^i_j \in TSub$; $M^i ob$ и $M^i sub$ – количество ассоциируемых с документом объективных и необъективных терминов, $M^i = M^i ob + M^i sub$.

Таким образом, процедура индексирования может быть представлена следующим отображением: $Ind: UD \to T$, где $T = TOb \cup TSub$.

Кроме того, процесс индексирования сопровождается формированием матрицы отношения между терминами и документами $W = \{w_{ij}\}_{N*(N_{ob}+N_{sub})}$. Значение элемента матрицы определяется как частота вхождения j-го термина в i-й документ (отношение числа вхождений термина к общему количеству слов в документе), взятая с коэффициентом, равным натуральному логарифму отношения общего количества документов к количеству документов, содержащих данный термин. Таким образом, элемент матрицы w_{ij} определяет вес j-го термина в i-м документе – чем больше частота термина в i-м документе и меньшее количество документов содержит j-й термин, тем больше вес j-го термина в i-м документе. Отсутствие термина в документе отражается равенством $w_{ij} = 0$.

На эффективность системы индексирования влияют два важных параметра. Полнота индексирования CI (Completeness of the Indexing) указывает степень распознавания тематики документа системой индексирования. Полная система индексирования по определению должна генерировать большое число терминов, отражающих все аспекты тематики документа; в случае неполной системы генерируется меньшее число терминов, соответствующих наиболее важным темам документа. Второй параметр – специфичность терминов ST (Specificity of Terms) – характеризует, насколько широк спектр охватываемых ими понятий. На выходе поиска по «широким» терминам (не обладающим ярко выраженной специфичностью) будет получено большое число полезных документов наряду со значительным количеством нерелевантной информации. Использование более специфичных («узких») терминов приведет к получению меньшего числа документов и может сопровождаться пропуском некоторой релевантной информации.

Процесс поиска происходит следующим образом. Пользователь формулирует свой запрос на естественном языке. Далее этот запрос транслируется в набор терминов, поэтому без потери общности можно принять, что на вход информационно-поисковой системы поступает вектор Q, состоящий из элементов множества T. Рассматриваемая гипотетическая ИПС осуществляет поиск, в результате которого формируется вектор $R = \{r^i\}_1^N$; значение r^i показывает, насколько документ i соответствует запросу пользователя. Далее пользователю выдается список документов $Rud = \{Ud^j \in Ud \mid r^j \geq \gamma\}$, упорядоченный по средневзвешенному коэффициенту соответствия. Величина $0 < \gamma \leq 1$ определяет порог коэффициентов соответствия документов, попадающих в результирующую выборку.

Математически, процесс поиска можно сформулировать следующим образом:

$$W \times Q = R. \qquad (1)$$

Вектор Q предварительно преобразуется в соответствие с матрицей W, чтобы можно было произвести умножение. На основе вектора R формируется список документов, выдаваемых пользователю.

Очевидно, что документ, содержащий термины запроса и имеющий средневзвешенный коэффициент, превосходящий пороговое значение γ, формально соответствует запросу. Однако такое совпадение не означает содержательного соответствия выданного документа запросу. Явление, при котором в ответ на запрос система выдает документы, не соответствующие запросу, называется поисковым шумом. В свою очередь, часть документов, релевантных запросу, может не попасть в результирующую выборку, тогда говорят о потерях информации. Информационный шум и потери информации могут быть выражены количественно с помощью коэффициентов полноты SC (Search Completeness) и точности поиска SP (Search Precision) [6, 17], являющихся показателями технической эффективности ИПС.

Под полнотой поиска понимается отношение числа найденных релевантных документов к общему числу релевантных документов в исследуемой совокупности. Под точностью поиска – отношение числа релевантных документов к общему числу полученных в результате поиска документов.

Пусть для заданного запроса Q некоторая операция *rel*() определяет подмножества релевантных документов $Ud_0 = \left\{ Ud^i \middle| rel(Ud^i, Q) = 1 \right\}$ и $Rud_0 = \left\{ Rud^i \middle| rel(Rud^i, Q) = 1 \right\}$ для всего множества документов и для результирующей выборки соответственно.

Тогда показатели полноты и точности поиска могут быть вычислены следующим образом:

$$SC = \frac{|Rud_0|}{|Ud_0|};\ SP = \frac{|Rud_0|}{|Rud|}. \qquad (2)$$

В идеальном случае значения обоих показателей должны приближаться к единице, но на самом деле приходится идти на компромисс. Индексирование по «широким» терминам обеспечивает высокие значения показателя полноты при низких показателях точности поиска, и наоборот, увеличение показателя точности поиска за счет использования «узких» терминов приводит к уменьшению значения полноты поиска. По этой причине эффективность многих ИПС оценивается значениями параметра точности при различных уровнях полноты.

Таким образом, математическая модель рассматриваемой гипотетической ИПС может быть представлена в виде совокупности:

$$ISS = \{UD, Ind, T, Q, R\}. \quad (3)$$

Формальное описание большинства современных информационно-поисковых систем соответствует сформулированной модели (3); процесс поиска информации в данных ИПС может быть описан с помощью (1).

Анализ модели задачи информационного поиска

Рассматриваемая модель (3) обладает рядом существенных недостатков. Один из них связан с достаточно сложной формализацией операции *rel*(). Как было отмечено выше, фактически существующие информационно-поисковой системы делают вывод только о формальной релевантности документа запросу пользователя на основании некоторого нормированного показателя соответствия. Если значение показателя превосходит заданный порог, то документ признается релевантным и попадает в результирующую выборку. Существуют различные методы определения данного показателя [6], обычно он вычисляется как средневзвешенная сумма индексов (терминов), взятых с весовыми коэффициентами. Однако решение о релевантности документа принимает в конечном итоге пользователь, инициирующий запрос.

Использование в качестве оценок эффективности ИПС показателей полноты и точности поиска (2) фактически сводит рассматриваемую задачу к задаче повышения качества индексирования. Индексирование всегда представляло одну из самых больших проблем при создании ИПС. Задача автоматического соотнесения с документами «высококачественных» терминов, хотя и изучается уже многие годы, остается пока нерешенной. Высококачественный поиск будет невозможен без решения этой сложной задачи.

Экспоненциальное расширение всемирной сети практически исключает создание и поддержание исчерпывающего индекса всего содержимого Интернет. Кроме того, увеличение размеров индекса приведет к значительному ухудшению эффективности поиска.

Также недостатком рассматриваемой модели можно считать практически полное отсутствие механизмов обратной связи ИПС с пользователем. Пользователь управляет только одним входным параметром модели (3), а именно вектором запросов в (1). Обычно поиск информации начинается с неточного и неполного запроса, который постепенно уточняется методом итераций. Тем самым, для повышения эффективности поиска пользователь должен уметь корректно сформулировать запрос на используемом в ИПС языке – ИПЯ (информационно-поисковом языке).

Повышение эффективности поиска

Недостатки современных информационно-поисковых систем, связанные с организацией процесса поиска, а также стремительное увеличение объемов размещенных в Интернет данных делают проблему повышения эффективности информационного поиска актуальной.

Учитывая то обстоятельство, что на сегодняшний день существующие глобальные ИПС практически единственное средство поиска информации в Интернет, для повышения эффективности поиска необходимо предложить способы, обеспечивающие выполнение следующих условий:

- максимально полное использование возможностей существующих ИПС;
- сведение к минимуму потерь, обусловленных присущими ИПС недостатками.

Как было отмечено выше, одновременно добиться высокого значения показателей полноты и точности поиска невозможно. Один из возможных вариантов решения данной проблемы заключается в формировании базы знаний по существующим ИПС, в которой бы описывались оптимальные пары значений показателей полноты и точности поиска для конкретной предметной области (например, «информатика», «мягкие вычисления»).

Уменьшение потерь информации в процессе поиска может быть достигнуто в результате увеличения точности передачи содержимого документов на ИПЯ [17] (что в конечном итоге приводит к проблеме повышения качества индексирования), а также в случае максимально полного соответствия запроса пользователя дескрипторам содержимого. Решение данной проблемы может быть найдено в использовании «интеллектуальных» посредников между пользователем и поисковой системой. Посредник должен обеспечивать перевод запроса пользователя на естественном языке на язык ИПС, осуществлять грамматические и синтаксические преобразования, осуществлять замену синонимов, ведение тезаурусов. Часть перечисленных функций уже реализована в современных ИПС.

С проблемой повышения качества индексирования взаимосвязана проблема релевантности результатов поиска. Каждая поисковая система использует свои алгоритмы определения релевантности документа, зачастую коэффициент релевантности для одного и того же документа может отличаться для различных ИПС. В данном случае, как одно из возможных решений, можно снова предложить «интеллектуального» посредника, который бы упорядочивал результаты запроса по иным критериям. Кроме того, необходимо учитывать и человеческий фактор – по некоторым исследованиям [3] большинство пользователей исследует только три первых результата работы поисковой системы.

В настоящее время высказываются предположения, что решение данных проблем может быть осуществлено в рамках двух сравнительно новых

направлений: нечеткой логики (Fuzzy Logic) и парадигмы мультиагентных систем (МАС) [7].

Применение нечеткой логики в задаче поиска информации в Интернет

Нечеткая логика [4] – направление научных исследований, начатых еще в работах Лотфи А. Заде, – является расширением классической (булевой) логики и основана на концепции "частичной правды". В некотором роде она служит методологическим расширением любой другой сколь угодно специфический теории, полученной путем "размывания" (fuzzification) ее базисных объектов, – предположении о том, что характеристическая функция может принимать любые значения на множестве [0; 1], а не только значения 0 либо 1.

В процессе информационного поиска в Интернет нечеткость проявляется в следующих моментах: пользователь, формулируя запрос, по сути, оперирует нечеткими понятиями (из-за полисемантичности слов); релевантность полученных результатов обычно выражается пользователем в нечетких оценках, зависящих от содержания найденных документов и предварительных оценок содержания.

В современных ИПС использование возможностей нечеткой логики обычно ограничивается расширением булевой модели поиска, но, как отмечается в [6], эффективность данного подхода остается достаточно низкой.

Наиболее очевидное применение методов нечеткой логики в рамках решения задачи информационного поиска – преобразование единиц естественного языка в термины ИПЯ. По мнению Лотфи А. Заде современные поисковые системы в дальнейшем должны превратиться в интеллектуальные системы, обладающие способностью выводить новые знания на основе имеющихся данных [18]. Основная проблема, как он полагает, заключается в особенности представления доступных знаний, а именно в их имманентной нечеткости. Качественное улучшение ИПС станет возможным только тогда, когда будет решена проблема интеграции информационных элементов [19]. В настоящее время предпринимаются попытки повысить эффективность поисковых систем с помощью разработки на основе методов нечеткой логики механизмов обратной связи с пользователем [20].

Применение МАС в задаче поиска информации в Интернет

Термин «мультиагентные системы» (МАС) используется для обозначения систем, состоящих из множества автономных сущностей – агентов [5]. Практически во всех работах, где дается определение того, что такое

агент и каковы его базисные свойства [8, 9], общим местом стало замечание об отсутствии единого мнения по этому поводу. Многообразие трактовок термина «агент» привело к тому, что в процессе разработки и реализации систем в рамках данного направления были предложены такие типы агентов, как автономные агенты, мобильные агенты, персональные ассистенты, интеллектуальные агенты, социальные агенты и т.д. Тем самым, вместо единственного определения базового агента, имеется множество определений производных типов. По аналогии с объектно-ориентированным программированием термин «агент» можно трактовать как виртуальный класс (совокупность виртуальных методов и свойств) [8], на основе которого порождаются классы-потомки – различные реализации объекта «агент».

Фактически, определение агента задается описанием его свойств. Обычно агент обладает набором из следующих свойств [5]:

- адаптивность – агент обладает способностью обучаться;
- автономность – агент обладает возможностью самостоятельных действий, формулируя для себя цели и выполняя действия для достижения поставленных целей;
- коллаборативность – агент может взаимодействовать с другими агентами несколькими способами, например, играя роль поставщика/потребителя информации или одновременно обе эти роли;
- коммуникативность – агенты могут общаться с другими агентами;
- способность к рассуждениям – агенты могут обладать частичными знаниями или механизмами вывода, например, знаниями, как приводить данные из различных источников к одному виду. Агенты могут специализироваться на конкретной предметной области;
- мобильность – способность к передаче кода агента с одного сервера на другой.

Существующие информационно-поисковые системы (например, Yandex, Rambler, Aport, Google – наиболее популярные ИСП в Рунете [11]) в контексте решаемой задачи поиска информации могут быть рассмотрены как прототипы агентов поиска. Они обладают свойствами автономности, коллаборативности, коммуникативности и мобильности. Агенты поиска, формируемые на основе одного из прототипов, должны также обладать свойством адаптивности и способностью к суждению.

В процессе разработки МАС необходимо решить следующие вопросы [5, 9]: Какой класс архитектуры использовать для определения агента? Что выбрать в качестве языка общения между агентами (включая выбор протокола передачи и формата сообщений)?

Наряду с агентами поиска в МАС должны быть реализованы агенты-координаторы (координации взаимодействий других агентов), агенты-интерфейсы (организация взаимодействия пользователя с МАС, организация взаимодействия МАС с агентами), агенты-эксперты (анализ предмет-

ной области, фильтрация полученных данных), агенты извлечения информации (извлечение данных, оценка и объяснение релевантности). Сама МАС должна обладать открытой архитектурой (должна позволять инкорпорировать в процесс поиска гетерогенные агенты), а также быть масштабируемой.

Таким образом, МАС информационного поиска должна обеспечивать выполнение следующих функций:

- координация работ ИПС;
- формулировка запросов с учетом специфики конкретной ИПС;
- обеспечение механизмов обратной связи с пользователем;
- оценка релевантности результатов;
- фильтрация информации;
- создание и хранение специализированных индексов.

Заключение

Интернет – огромное хранилище распределенных оцифрованных данных, для использования его потенциала необходимо наличие действенных инструментов поиска и обработки данных. Именно низкая эффективность современных средств поиска информации является на сегодняшний день главной проблемой на пути превращения Интернет в интеллектуальную сеть. В рамках решения данной проблемы предлагаются различные методы и подходы [12], разрабатываются прототипы «интеллектуальных» поисковых систем [21]. Объединяющим фактором в этих попытках решения проблемы является использование концепции агентов и мультиагентных систем, применение методов нечеткой логики и технологий распределенных вычислений.

Литература

1. ISC Internet Domain Survey // Internet Systems Consortium, Inc. http://www.isc.org/index.pl?/ops/ds/reports/2004-01/

2. Computer Industry Almanac, Press Releases. - 2002, March 21. http://c-i-a.com/pr032102.htm

3. Jansen J. Impatient Web Searchers Measure Web Sites' Appeal In Seconds // Penn State Live. – 2003. – June 25. http://live.psu.edu/index.php?cmd=vs&story=3364

4. Brule J.F. Fuzzy systems tutorial. http://life.anu.edu.au/complex_systems/fuzzy.html

5. Franklin S. and Graesser A. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents // The Third International Workshop on Agent Theories, Architectures, and Languages. - Springer-Verlag, 1996. http://www.dfki.uni-sb.de/~jpm/atal96.html

6. Gudivada V.N. Information search on World Wide Web // Computer Weekly. - 1997. - N 35. - P. 19- 21, 26, 27.

7. Guifoyle C., Warner E. Intelligent Agents: the New Revolution in Software // Ovum. IBM Agents, 1994. http://activist.gpl.ibm.com:81/WhitePaper/ptc2.htm

8. Nwana H.S., Ndumu D.T. An Introduction to Agent Technology // BT Technology Journal. - 1996. - V. 14. - N 4. - P. 55-67

9. Wooldridge M., Jennings R. Intelligent Agents: Theory and Practice // Knowledge Engineering Review. - 1995. - V. 10. - N 2. - P. 115-152

10. Касумов В.А. Поисковые механизмы библиотечно-информационных систем Internet. http://library.ntu-kpi.kiev.ua/html/arhiv/arh177/tom1/444/Doc15.HTML

11. Королев И. Поисковые системы: день сегодняшний // Рунет.ру. - 2003. http://www.runet.ru/analitika/4157.html

12. Морозов А.А. Об одном подходе к логическому программированию интеллектуальных агентов для поиска и распознавания информации в Интернет // Журнал радиоэлектроники. - 2003. http://jre.cplire.ru/jre/nov03/1/text.html

13. Тихонов В. Поисковые системы в сети Интернет. http://www.citforum.ru/internet/search/searchsystems.shtml

14. Храмцов П. Информационно-поисковые системы Internet // Открытые системы. - 1996. - № 3. http://www.osp.ru/cw/1996/03/46.htm

15. Храмцов П. Моделирование и анализ работы информационно-поисковых систем Internet // Открытые системы. - 1996. - № 6. http://www.osp.ru/cw/1996/06/46.htm

16. Храмцов П. Поиск и навигация в Internet // Открытые системы. - 1996. - № 20. http://www.osp.ru/cw/1996/20/31.htm

17. Чурсин Н. Популярная информатика. - Киев: Техника, 1982.

18. Zadeh L.A. From Search Engines to Question-Answering Systems // «2003 BISC FLINT-CIBI International joint workshop on Soft Computing for INTERNET and Bioinformatics». http://www-bisc.eecs.berkeley.edu/FLINTCIBI/abstracts.html

19. Korotkikh G. Fuzzy Spectral Patterns for Information Integration in a Search Engine // «2003 BISC FLINT-CIBI International joint workshop on Soft Computing for INTERNET and Bioinformatics». http://www-bisc.eecs.berkeley.edu/FLINTCIBI/abstracts.html

20. John R.I. The Role of User Modeling in Information Retrieval from the WWW // «2003 BISC FLINT-CIBI International joint workshop on Soft Computing for INTERNET and Bioinformatics». http://www-bisc.eecs.berkeley.edu/FLINTCIBI/abstracts.html

21. Tang Y. and Zhang Y. Genetic Fuzzy Neural Agents Using Type-2 Fuzzy Reasoning for Intelligent Web Information Search Task // «2003 BISC FLINT-CIBI International joint workshop on Soft Computing for INTERNET and Bioinformatics». http://www-bisc.eecs.berkeley.edu/FLINTCIBI/abstracts.html