Математические структуры и моделирование
2001, вып. 8, с. 126–131

УДК 681.518

О КОНЦЕПЦИИ ИНДЕКСАЦИИ ИНФОРМАЦИОННЫХ РЕСУРСОВ СЕТИ ИНТЕРНЕТ

И.А. Земсков

In this article two methods building a full-text index for the Web are described.

Введение

Настоящая статья посвящена обсуждению путей повышения эффективности поисковых средств сети Интернет. В ней делается попытка обобщить идеи и соображения, возникшие у автора на основе личного опыта работы с поисковыми средствами сети Интернет, а также на основе многочисленных публикаций на эту тему.

В качестве объекта исследования выбрана система индексирования информационных ресурсов в Интернет или т.н. индексирующий робот. Автор сознательно ограничивает область рассмотрения, вычленяя из всех проблем, связанных с поиском информации в Интернет, проблему индексации, так как, на его взгляд, эта проблема была незаслуженно лишена внимания исследователей.

Целью настоящей статьи является анализ функционирования индексирующих роботов поисковых систем. Это позволит вплотную подойти к вопросу создания математических моделей систем индексирования информационных ресурсов и прогноза дальнейших условий их функционирования. Также целью статьи является описание возможных проблем при технической реализации предлагаемых нововведений.

1. Описание процесса индексации

Концепция работы индексирующих роботов, а их насчитывается уже порядка 270 [1], не менялась со времён изобретения первых поисковых систем Интернет. До недавнего времени подобный факт можно было оставлять без внимания, но бурный рост сети Интернет вносит свои коррективы в расстановку приоритетов решаемых задач. Согласно [2] число серверов World Wide Web, а значит, и документов в этой распределенной информационной системе Интернет удваивается каждые 60 дней. Однако происходит не только удвоение количества ресурсов, но и изменение уже существующих. А это означает, что такие системы, как

E-mail: zemskov@univer.omsk.su
Омский государственный университет

Altavista и Lycos, обязаны не только обновлять свои поисковые индексы непрерывно, но и должны постоянно увеличивать зону охвата индексируемых ресурсов. В настоящее время практически все крупные поисковые системы этого типа как в России, так и за границей имеют своих индексирующих роботов, которые работают постоянно и всё же не поспевают за стремительным развитием обрабатываемой среды. Следует отметить, что время «повторной» индексации ресурса с каждым разом увеличивается. Это, в свою очередь, может приводить к «утере» адекватного отображения реального состояния ресурсов.

В современной концепции процесса индексации мы можем выделить два типа узлов сети Интернет, принимающих непосредственное участие в работе:

узел-донор — на нём расположена коллекция веб-страниц, предназначенная для индексации.

узел-обработчик — на нём происходит три процесса: первоначальная обработка скачанных с узла-донора веб-страниц; процесс построения индекса и его сохранения; выполнение поисковых запросов на основе полученного индекса.

Отбрасывая процесс выполнения поискового запроса, можно сказать, что процесс индексации протекает в две стадии: во-первых, узел-обработчик «скачивает» с узла-донора его коллекцию веб-страниц, а во-вторых, на узле-обработчике происходит обработка полученной коллекции для получения т.н. поискового массива (индекса).

Основное внимание исследователей направлено на изучение и оптимизацию процессов, происходящих на узле-обработчике. Например, в [3] рассматриваются четыре наиболее популярные меры близости, используемые в информационно-поисковых системах Интернет для ранжирования найденных документов. Большая работа по оптимизации протекающих процессов была проделана в работе [4]. Её авторы предложили разделить узел-разработчик на два узла: узел-индексатор и узел-поисковик.

Однако, независимо от разработчиков, неизменными остаются два момента в процессе индексации: во-первых, для того чтобы начать строить индекс, нужно «скачать» коллекцию веб-страниц на узел-обработчик; во-вторых, начальным этапом обработки становится избавление от т.н. «мусора» в виде HTML-кода (что считать «мусором», зависит от алгоритма построения индекса, но как минимум это HTML-код).

Акцентирование внимания на этих моментах приводит к формулированию целого ряда важных вопросов:

1. Какой объём «мусора» скачивает к себе робот?
2. Сколько времени и других собственных ресурсов тратит узел-обработчик на обработку «мусора»?
3. Как посчитать то количество системных ресурсов (например: вычислительная мощность процессоров, ОЗУ, дисковое пространство, ширина канала связи с Интернет), которое потребуется узлу-обработчику для поддержания в актуальном состоянии своего поискового массива, построенного по коллекциям веб-страниц максимально возможного числа узлов-доноров сети Интернет?

Из поставленных вопросов становится очевидным тот факт, что узел-обра-

ботчик становится источником бесполезной нагрузки на сеть Интернет, т.к. некоторая часть (её размер ещё предстоит выяснить) его работы является обработкой «мусора», который за ненадобностью впоследствии выбрасывается.

2. Новая концепция индексации

Совсем избавиться от затрат на обработку «мусора» мы не в состоянии, т.к. он служит нам той «оболочкой» для информации, которая делает Интернет Интернетом. Но нам вполне по силам взаимовыгодно разделить эти затраты между всеми членами Интернет-сообщества (имеются в виду только владельцы узлов-доноров и узлов-обработчиков). Или, другими словами, с учётом предыдущих рассуждений становится очевидным желание перераспределить функции, выполняемые в процессе индексации, между узлами-донорами и узлом-обработчиком.

Опишем узлы, участвующие в процессе, с учётом новых функций:

узел-донор — на нём, помимо коллекции веб-страниц, расположен модуль-робот, отвечающий за выполнение предварительной обработки веб-страниц и последующей передачи результатов обработки на узел-обработчик для построения поискового массива;

узел-обработчик — на нём происходит три процесса: приём обработанных на узле-доноре веб-страниц, фактически приём «чистой» информации, лишённой «мусора»; процесс построения индекса и его сохранения; выполнение поисковых запросов на основе полученного индекса.

Рассмотрим подробнее процессы, происходящие на узле-доноре:

1. Процесс инициализации — происходит в начальный момент функционирования самой коллекции веб-страниц.
2. Обнаружение «цели» для обработки — под целью понимается веб-страница. Может быть два вида целей: веб-страница ранее не была обработана для последующей передачи узлу-обработчику и веб-страница претерпела изменение содержимого, но не изменила своих «координат». Для осуществления этой функции может потребоваться организация локального индекса обработанных веб-страниц. Стоит заметить, что в момент процесса инициализации локальный индекс пуст. Техническая реализация процесса обнаружения «цели» может быть основана на двух моделях поведения: пассивной или активной. Пассивная модель заключается в простом «прослушивании» ответов на внешние запросы к коллекции веб-страниц и нахождении т.н. «цели» для обработки. В защиту этой модели поведения можно сказать то, что таким образом мы проиндексируем всю действительно «интересную» информацию, располагающуюся на узле (так как, если к ней обратились, значит она интересна). Активная модель поведения базируется на алгоритмах поведения существующих индексирующих роботов. Данная модель, по мнению автора, имеет больше минусов, чем плюсов. Например, из основных минусов можно выделить следующий: она создаёт неоправданно большую дополнительную вычислительную нагрузку на системные ресурсы узла-донора благодаря своим сложным эвристическим

алгоритмам нахождения «цели» обработки в коллекции веб-страниц, которые потребуется постоянно обновлять в связи с развитием технологий, поддерживаемых при создания веб-страниц.

3. Обработка найденной «цели» – обработка может заключатся как в простом избавлении веб-страницы от определённого набора HTML-тегов, так и в более детальной проработке, в зависимости от последующих потребностей индексирующего узла-обработчика.
4. Передача на узел-обработчик локального индекса, всего или только некоторой части. Другими словами, на индексирующий узел-обработчик должна передаваться не вся веб-страница, а только её информационное наполнение, представленное в нужном для индексации виде, плюс служебная информация (например, помимо информационного наполнения страницы, индексирующему процессу нужны «координаты» веб-страницы).

Дополнительное пояснение требуется термину «локальный индекс»: роль локального индекса может выполнять как минимум локальная (для узла-донора) коллекция обработанных веб-страниц вместе с сопоставленными им координатами реальных веб-страниц. Под координатами веб-страниц может пониматься URL адрес веб-страницы. Существование локального индекса видится принципиально важным, так как, во-первых, с помощью него удаётся принять решение о нахождении новой веб-страницы в коллекции узла-донора, избегая при этом обращения к узлу-обработчику за дополнительной информацией; во-вторых, становится возможным (практически мгновенно) отследить изменение в уже проиндексированных страницах и сообщить о характере изменений на узел-обработчик. Тем самым становится возможным изменить стандартное значение термина «повторная индексация», т.к. теперь под ним будет пониматься не повторная обработка всей коллекции веб-страниц, а выборочная обработка новых или изменившихся веб-страниц коллекции узла-донора.

Согласие с целесообразностью организации локального индекса на основе этих предположений влечёт за собой возникновение вопроса о его размере, т.е. вопроса о том количестве ресурсов узла-донора, которое нужно выделить на поддержку локального индекса. И хотя количество и тип ресурсов ещё предстоит изучить, но уже сейчас можно предположить одно из возможных направлений минимизации дискового пространства, занимаемого индексом. Например, на узле-доноре достаточно хранить не полную версию веб-страницы после её обработки, а только некий код, полученный на основе избавленного от «мусора» содержимого веб-страницы, однозначно идентифицирующий её содержимое. Вполне достаточным может оказаться использование кодирования «в одну сторону» и последующего сравнения получившегося кода с уже имеющимся для этой веб-страницы. Однако нужно иметь в виду, что кодирование должно заметным образом уменьшать размер обрабатываемой информации, т.к. иначе его использование лишено смысла.

Рассмотрение вопроса организации локального индекса, с точки зрения программной реализации, наталкивается на ряд вопросов по обеспечению безопасности (или, другими словами, ограничения доступа к закрытой информации) конфиденциальной информации, возможно находящейся в коллекции веб-стра-

ниц.

Немного слов скажем о передаче локального индекса узла-донора на узел-обработчик. В новой ситуации, вызванной перемещением функций первоначальной обработки веб-страниц на узел-донор, узел-обработчик сохранил только две ресурсоёмкие функции — это построение поискового массива (индекса) и выполнение поисковых запросов. Но взамен утерянной функции первоначальной обработки он приобрёл функцию приёма локальных индексов узлов-доноров. Сама по себе функция не является источником большой вычислительной нагрузки, но в условиях большого количества узлов-доноров она может стать серьёзной проблемой организации стабильной работы узла-обработчика. Можно предложить несколько подходов к решению обозначенной проблемы. Первый подход заключается в том, чтобы описать механизм «отложенной передачи локального индекса», который по своей сути является ожиданием освобождения ресурсов узла-обработчика в случае их «занятости». Второй подход реализуется выделением необходимого количества системных ресурсов на основе прогноза возможной максимальной загруженности узла-обработчика. Оба подхода можно изучить с помощью соответствующей имитационной модели очереди с отказом в обслуживании.

С точки зрения технической реализации процесса взаимодействия узла-донора и узла-обработчика, будет интересным получить ответы на следующие вопросы:

1. Какой из двух узлов должен стать инициатором взаимодействия? Узел-донор до сих пор был пассивным участником, но теперь имеет смысл рассмотреть возможность его активизации, т.к. только ему известен момент актуального обновления локального индекса.
2. Возможно ли обойтись в реализации процесса взаимодействия только средствами протокола HTTP или HTTPS? Ответ на этот вопрос во многом будет зависеть от ответа на предыдущий вопрос.
3. Какие меры нужно предпринять, чтобы исключить возможность проведения любых несанкционированных действий по отношению к обоим узлам? В вопросе кроется большая проблема защиты рассматриваемой системы от хакерских атак. Решение этой проблемы во многом будет влиять на получение путёвки в жизнь предлагаемой технологии индексации.

Заключение

Многие направления будущих исследований формулировались по ходу статьи, но первоочередную цель сформулируем ещё раз:

> чтобы подтвердить или опровергнуть выгоду от реализации выдвинутой гипотезы об экономической целесообразности перенесения части индексирующего робота на узел-донор, следует создать имитационную модель взаимодействия узлов-доноров и узла-обработчика по старой концепции индексации, а также имитационную модель взаимодействия узлов-доноров и узла-обработчика по новой концепции индексации, описанной в этой статье.

Литература

1. The Web Robots Pages. – `http://www.robotstxt.org/wc/active/html/index.html`
2. Попов И.И., Храмцов П.Б. *Распределение частоты встречаемости терминов для линейной модели информационного потока* // НТИ. 1991. Сер.2. №2. С.23-26.
3. Budi Yuwono, Dik L.Lee. *Search and Ranking Algorithms for Locating Resources on the World Wide Web* // In Proceedings of the Forth International Conference on the World Wide Web. New York. November. 1995.
4. Sergey Melnik, Sriram Raghavan, Beverly Yang, and Hector Garcia-Molina *Building a Distributed Full-Text Index for the Web.* – `http://www-diglib.stanford.edu/cgi-bin/get/SIDL-WP-2000-0140`
5. Храмцов П.Б. *Моделирование и анализ работы информационно - поисковых систем Internet* // Открытые Системы. 1996. №6.