Информационный агент для формирования тематической коллекции электронных документов

Рушди А. Амамра
Санкт-Петербургский Государственный Технический Университет
rushbeth@usa.net

В данной работе описывается архитектура информационного агента, предназначенного для формирования тематических коллекций электронных документов. К основным особенностям данного подхода относятся:

- Использование метода вероятностного латентного семантического индексирования для выявления основных тем, затрагиваемых в формируемой коллекции.
- Использование двух типов фильтров для документов, рекомендуемых для включения в коллекцию: фильтр, построенный на основе анализа содержимого коллекции, и фильтр, построенный на основе анализа архива пользовательских запросов к данной коллекции.

Введение

В настоящее время наиболее широко используемые поисковые системы для поиска информации в Интернет (Alta Vista, Google и другие) основаны на централизованной архитектуре, при которой одна организация ответственна за индексирование всей опубликованной в Интернет информации. Такое решение проблемы поиска информации нельзя признать масштабируемым, о чем и говорят данные, показывающие, что доля индексированного Интернет ежегодно падает.

Масштабируемым решением может являться система распределенного поиска с децентрализованной архитектурой (см., например, [8]). В системах такого типа различными независимыми организациями индексируется относительно небольшие тематически специфические области, а запросы пользователей автоматически перенаправляются в те тематические коллекции (тематические индексы), тематика которых в наибольшей степени соответствует тематике запроса.

Важнейшая задача автоматического (полуавтоматического) формирования тематических коллекций возлагается на информационного агента (сетевого робота, кроулера). Информационные агенты для работы в WWW рассматривались в большом числе работ [6],[1], [7], [3].

Задача агента, предлагаемого в данной работе, состоит в пополнении коллекции новыми релевантными ее тематике документами. Как правило, работа такого агента начинается с некоторого множества HTML документов (ядра коллекции), заданных администратором коллекции. Далее агент загружает документы, на которые ссылаются уже включенные в коллекцию документы, и рекомендует к включению в коллекцию те из них, которые проходят через фильтр, сформированный на основе анализа ядра коллекции. Однако такой подход имеет свои недостатки. Например, ядро коллекции может не отражать некоторые важные темы (особенно недавно появившиеся) и, следовательно, агент будет пропускать документы, которые следовало бы включить в коллекцию. В данной работе предлагается при фильтрации документов использовать два фильтра и рекомендовать в коллекцию документ, прошедший хотя бы через один из них. Первый фильтр отражает содержимого ядра коллекции и представляет тем самым интересы ее администратора. Второй основан на анализе архива запросов пользователей, полученных данной коллекцией за определенный период времени. Запросы пользователей отражают информационные

потребности сообщества пользователей, которые должны учитываться при пополнении коллекции новыми документами. Это позволит поддерживать в течении всего времени жизни коллекции ее адекватность интересам пользователей.

Построение обоих фильтров основано на использовании относительно нового метода вероятностного латентного семантического индексирования (PLSI) [5]. Использование данного метода позволяет выявить в коллекции заданное число латентных факторов, представляющих более узкие подтемы, затронутые в документах коллекции. Для каждого фактора можно оценить силу связи данного фактора с каждым словом из словаря коллекции и с каждым ее документом. Все это дает возможность выявить для данной коллекции наиболее важные темы и наиболее важные для каждой темы слова, которые и формируют фильтр, основанный на анализе ядра коллекции. Заметим, что сам метод латентного семантического индексирования был предложен значительно ранее в работе [4] и далее развивался в ряде работ (см., например, [2]). При этом для выделения факторов строилась малоранговая аппроксимация матрицы, отражающей связи между документами и словами.

При построении фильтра, основанного на анализе архива запросов, используются только те слова, которые встречаются в запросах. Веса слов, которые входят в словарь коллекции, определяются на основе результатов применения PLSI. Для оценивания весов слов, не входящих в словарь коллекции, предлагается новый подход, учитывающий семантическую близость слов, входящих в один относительно короткий запрос.

Фильтрация новых документов

Новый документ, загруженный из Интернет, рекомендуется коллекции, если он проходит хотя бы через один из фильтров : фильтр коллекции и фильтр запросов.

При построении фильтра коллекции выполняется анализ ядра коллекции методом PLSI. Обозначим через $z_i,\ i = 1,...,n$ так называемые скрытые (латентные) факторы - идентификаторы относительно узких тем, представленных в ядре коллекции. Метод PLSI позволяет вычислить оценки для следующих вероятностей:

$P(z_i)$ - вероятность того, что случайно выбранный из ядра коллекции документ относится к тематике z_i,

$P(d_j \mid z_i)$ - вероятность того, что документ d_j относится к тематике z_i,

$P(w_j \mid z_i)$ - вероятность того, что слово w_j относится к тематике z_i.

Для всех слов из словаря коллекции (слова, встречающиеся в ее ядре) вычислялись веса по формуле

$$\mathrm{W}(\mathrm{w}_\mathrm{j}) = \sum_{i=1,...,k} P(z_i)P(w_j \mid z_i)$$

В качестве фильтра коллекции выбираются заданное количество слов из словаря коллекции с наибольшими весами.

Фильтр запросов отражает информационные потребности всего сообщества пользователей. В этот фильтр входят только те слова, которые встречаются в архиве запросов. Среди них имеются как слова из

словаря коллекции, так и новые слова, представляющие новые темы, не представленные в ядре коллекции. Веса слов из словаря коллекции уже вычисленны при построении фильтра коллекции. Для вычисления весов новых слов используется следующий подход.

Построим граф G, вершинами которого являются все слова, встречающиеся в архиве запросов. Две вершины w_i и w_j соединены ребром, если эти два слова встречается вместе хотя бы в одном запросе. При построении оценки веса нового слова исходим из следующего принципа: вес нового слова равен среднему арифметическому весов всех слов, являющихся соседями данного слова в графе G Для вычисления этих весов используется метод простой итерации. Сходимость метода гарантируется, если каждое новое слово встречается с каким-либо из слов из словаря коллекции хотя бы в одном запросе из архива запросов. В противном случае итерационный процесс прерывается по достижении заданного числа итераций. При этом веса некоторых новых слов могут быть оценены с большой погрешностью, но эта погрешность будет уменьшаться при появлении новых запросов, включающих данные новые слова.

В фильтр запросов включаются заданное число слов (встречающихся в архиве запросов) с наибольшими весами.

При сопоставлении нового документа с фильтром вычисляется скалярное произведение tf - профайла документа с фильтром. Документ проходит через фильтр, если это скалярное произведение превышает заданный порог.

Основной алгоритм

Алгоритм функционирования агента содержит следующие основные этапы

1. Генерация фильтра коллекции
   Фильтр коллекции строится на основе анализа содержимого ядра коллекции в соответствии с описанием приведенным в предыдущем пункте

2. Генерация фильтра запросов
   Фильтр запросов строится на основе анализа архива запросов в соответствии с описанием приведенным в предыдущем пункте

3. Инициализация дерева URL
   Качество работы агента определяется как качеством используемых фильтров, так и организацией очереди URL документов, подлежащих загрузке. В рассматриваемом алгоритме формируется дерево URL, которое и используется при формировании очереди. На этапе инициализации формируется дерево с двумя уровнями. На первом уровне - корень дерева, на втором - узлы, содержащие стартовые URL заданные администратором коллекции. Каждому узлу v приписывается оценка $P(v)$ вероятности того, что ссылка из соответствующего документа указывает на документ

релевантный тематике коллекции. Для стартовых URL на этапе инициализации эти оценки принимаются равными 1.

4. Выбор URL (этот пункт и последующие повторяются в течении всего времени работы агента) В дереве URL выбирается узел v , для которого $P(v)$ максимально (при этом не выбираются URL, указывающие на недавно посещенные сайты). Если документ с соответствующим URL еще не загружен, то для последующей загрузки выбирается данный URL. В противном случае случайным образом выбирается еще не рассмотренная ссылка из этого документа на новый документ в формате html. Если не рассмотренных ссылок нет, то выбирается другой узел.

5. Загрузка и фильтрация документа
   С помощью программы wget загружается документ с заданным URL. Выполняется разбор текста документа - выделяются ссылки на другие html документы и вычисляется tf -профайл загруженного документа. Фильтрация документа описана в предыдущем разделе.

6. Модификация дерева URL
   Если документ не прошел фильтрацию, то он не рекомендуется для включения в коллекцию. Если для URL данного документа в дереве URL уже имеется узел, то он помечается как нерелевантный и оценка вероятности релевантности исходящих из него ссылок принимается равной нулю. В противном случае оценка $P(v)$ уменьшается. Новое значение равно

$$P(v) = \frac{1 + links_+(v)}{1 + links_+(v) + links_-(v)}$$

где $links_+(v)$ равно числу проверенных релевантных ссылок из документа в узле v , а $links_-(v)$ - числу проверенных нерелевантных ссылок из того же документа.

Если документ прошел фильтрацию, то он рекомендуется для включения в коллекцию. Если для URL данного документа в дереве URL уже имеется узел, то он помечается как релевантный и оценка вероятности релевантности исходящих из него ссылок принимается равной 1. В противном случае оценка $P(v)$ увеличивается (если оно было меньше 1). Новое значение равно

$$P(v) = \frac{1 + links_+(v)}{1 + links_+(v) + links_-(v)}$$

где $links_+(v)$ равно числу проверенных релевантных ссылок из документа в узле v , а $links_-(v)$ -числу проверенных нерелевантных ссылок из того же документа. Кроме того, формируется новый узел w , содержащий URL загруженного документа. Величина $P(v)$ принимается равной 1. В дереве URL узел w является сыном узла v .

Методика проведения экспериментов

В качестве ядра коллекции была выбрана небольшая коллекция по тематике информационного поиска. В качестве архива запросов использовались 100 относительно коротких фраз, выбранных из других документов, связанных с тематикой информационного поиска. Цель экспериментов состояла в выборе оптимальных значений порогов, используемых в фильтре коллекции и в фильтре запросов. Были выбраны 25 пар значений пороговых величин. Для каждой пары агент стартовал с одного и того же множества стартовых URL и останавливался после загрузки $n_+ = 200$ прошедших фильтрацию документов. Эксперт оценивал реальный процент релевантных документов среди всех документов прошедших фильтрацию - p. В качестве итоговой оценки точности работы агента для данной пары пороговых значений принимается величина

$n_+ p / n_{all}$, где n_{all} - число всех загруженных документов. В результате для лучшей пары пороговых величин точность работы агента оценивается величиной 0.793. Заметим, что пости все документы, рекомендованные агентом к включению в коллекцию, были признаны экспертом релевантными ее тематике.

Заключение

В данной работе описывается архитектура информационного агента, предназначенного для формирования тематических коллекций электронных документов. Метод вероятностного латентного семантического индексирования использовался для формирования фильтра коллекции, построение которого было основано на анализе содержания ядра коллекции. Таким образом, фильтр коллекции отражает тематику коллекции с точки зрения ее администратора. Дополнительно был использован фильтр запросов, при построении которого проводится анализ архива запросов пользователей, полученных данной коллекцией. Эти запросы отражают информационные потребноти сообщества пользователей, и их учет позволяет пополнять коллекцию документами, релевантными сегодняшним потребностям большинства пользователей.

Результаты экспериментов показали достаточно высокую точность поиска, обеспечиваемую предложенным агентом. Это, в частности, иллюстрирует эффективность использования вероятностного латентного семантического индексирования в задачах информационного поиска.

References

1- Robert Armstrong, Dayne Freitag, Thorsten Joachims, and Tom Mitchell. Webwatcher: A learning apprentice for the world wide web. In *AAAI Spring Symposium on Information Gathering*, pages 6-12, 1995.

2- Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. Technical Report UT-CS-94-270, 1994.

3- Daniel Boley, Maria Gini, Robert Gross, Eui-Hong (Sam) Han, Kyle Hastings, George Karypis, Vipin Kumar, Bamshad Mobasher, , and Jerome Moor. Document categorization and query generation on the world wide web using webace. *AI Review*, 13(5-6):365-391, 1999.

4- Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391-407, 1990.

5- Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. of the SIGIR'99*, pages 50-57, Berkley, CA, USA, August 1999.

6- Henry Lieberman. Letizia: An agent that assists web browsing. In Chris S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 924-929. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1995.

7- Martin E. Meller. Machine learning based user modeling for www search , http://citeseer.nj.nec.com.

8- A. Patel, L. Petrosjan, and W. Rosenstiel, editors. *OASIS: Distributed Search System in the Internet*. St. Petersburg State University Published Press, St. Petersburg, 1999.