КОМП'ЮТЕРНІ ЗАСОБИ, МЕРЕЖІ ТА СИСТЕМИ

УДК 681.3

В.Г. ПИСАРЕНКО, Л.С. ХАРЧЕНКО,
Н.Ф. ГОРИНА

СРАВНИТЕЛЬНАЯ ОЦЕНКА РЕЛЕВАНТНОСТИ ПОЛУЧАЕМЫХ ДАННЫХ ДЛЯ ПРОБЛЕМНО-ОРИЕНТИРОВАННЫХ ТЕХНОЛОГИЙ УСКОРЕННОГО ПОИСКА В КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ ХРАНИЛИЩАХ

На основании анализа существующих методов ускоренной оценки сравнительной релевантности анализируемых документов, предлагается шкала предпочтительности использования различных комбинаций подобных методов в зависимости от характера анализируемого документа и поставленных пользователем целей. Предлагаемая шкала предпочтительности актуальна для проблемно-ориентированных технологий ускоренного поиска данных при формировании и эксплуатации корпоративных информационных хранилищ.

Учитывая актуализацию использования интеллектуальных технологий при сборе, хранении, обработке данных о производственной деятельности крупной корпорации разрабатываются технологии формирования корпоративных информационных хранилищ (ИХ) данных с методиками быстрого поиска (Data Mining) и очистки данных (data cleaning, data cleansing или scrubbing) для последующего использования в ведомственной информационно-аналитической системе (ИАС) поддержки принятия решений руководством корпорации. В современных условиях объемы хранимой информации могут достигать от сотен гигабайт до нескольких терабайт (например, хранилища, реализованные в технологии Storage Works Raid Array и при использовании мощных серверов семейства Sun Enterprise, показавших рекордную пропускную способность). При этом для таких массивов данных необходимо обеспечить быстрый поиск необходимых данных для получения ответов на SQL-запросы пользователей системы разных категорий в ограниченные сроки, включая режим, близкий к online.

При формировании корпоративного информационного хранилища на первом этапе реализуется технология сбора первичной (неверифицированной) информации по заданно-

му критерию с помощью некоторой исполняемой программы поиска (*i*-я поисковая программа) в соответствии со схемой, изображенной на рисунке.

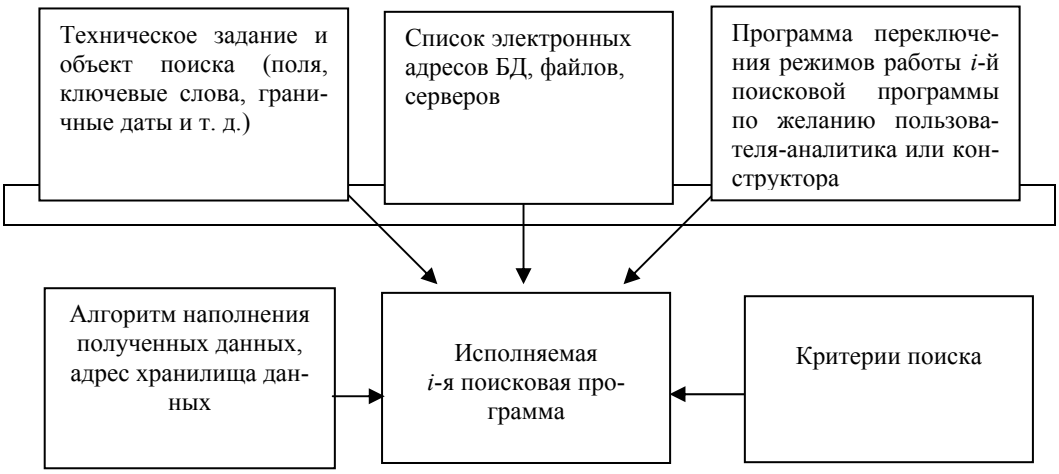

РИСУНОК. Технология сбора первичной информации по заданному критерию для информационного хранилища

На втором этапе количественно оценивается эффективность для пользователя собранной в ИХ первичной информации, в соответствии с чем данные очищаются, верифицируются и отбираются для ИХ верифицированных данных.

Для целей сравнительной количественной оценки эффективности для пользователя (релевантности) различных конкурирующих технологий быстрого поиска обычно используют лингво-статистические методы (ЛСМ) и семантические методы (СМ). Для определения фактической релевантности лингво-статистическим методом предлагается использовать комплексный критерий

$$R_{\Sigma} = \sum_{i=1}^{N} \sigma_i \cdot K_i ,$$

где N – количество критериев определения релевантности, σ_i – коэффициент значимости *i*-го критерия, K_i – количественное значение *i*-го критерия для анализируемого документа. В качестве критериев K_i предлагается использовать следующие критерии: количество ключевых слов (КС) запроса в документе; удельный вес ключевых слов, т. е. отношение количества КС к общему количеству слов в документе; индекс цитируемости; время существования данного документа (на сайте сети).

Из семантических методов целесообразно использовать метод латентного семантического анализа (ЛСА). В методах ЛСА рассматриваются документы или части документа как набор семантически самостоятельных единиц. Согласно этому подходу система ЛСА преследует цель установить степень смыслового соответствия данного документа некоторому априорно сформулированному

пользователем перед поиском «идеальному» образу документа. Происходит это в соответствии с принципами построения векторных пространств смысловых связей и представляет скорее математическую, чем лингвистическую задачу. Эта технология основана на формальной математической модели извлечения информации из текстов и предполагает, что каждому документу, как и любой законченной по смыслу единице (терму) текстовой информации (например, главы, абзацы, предложения), можно поставить в соответствие некоторый вектор в пространстве смысловых связей между этими термами. Главная идея такого подхода основана на предположении, что смысловая составляющая документа может быть представлена как совокупность терминов-понятий, встречающихся с разной частотой в этом документе. Согласно этой гипотезе каждый терм можно представить как вектор [1]:

$$d = (t_1, t_2, \ldots, t_n),$$

где t_i, $1 \le i \le n$ – это неотрицательное значение, отражающее частоту появления i-го терма в документе d.

Каждый вектор, представляющий какой-либо документ, задает местоположение последнего в семантическом пространстве термов, описывающих предметную область, к которой относятся сами эти понятия, и документы, в которых они используются. О семантической близости тех или иных документов судят по расстояниям между векторами, составляющими пространство термов-документов.

ЛСА дает следующие преимущества:

- решается проблема полисемии. Суть этой проблемы в том, что один и тот же термин может иметь различные значения в том или ином контексте. При этом решение проблемы полисемии в ЛСА в поиске не простого соответствия терминов в различных документах, а анализируется сам контекст в целом, со всеми его смысловыми связями и отношениями [2];

- решается проблема синонимии. Термин может вообще не встречаться в релевантных (близких по смыслу) документах и, тем не менее, окажет свое «скрытое», латентное влияние на степень близости текстов. Другими словами, проблема может быть описана с помощью различной терминологии, и это не помешает системе ЛСА определить семантическую близость всех подобных описаний [3].

Остановимся подробнее на основных особенностях ЛСА, выделяющих его среди множества методов извлечения информации и основанных как на лингвистических, так и на статистических данных, получаемых при анализе текстов.

Анализ показывает, что ЛСА на 20-30% эффективнее существующих методов [4]. Так, ЛСА использует действительно многомерное представление информации, поскольку сопоставляет выбранной ключевой фразе из N слов соответствующий N-мерный вектор, каждая компонента которого отвечает апостериорной встречаемости соответствующего элемента ключевой фразы в i-ом документе, $i = 1, 2, \ldots, M$. Таким образом, все термины и документы, подвергаю-

щиеся анализу, представляются в виде матрицы M × N. Кроме того, ЛСА изначально ориентирован на анализ крупных текстовых массивов, что действительно может найти применение при анализе релевантности объёмных документов.

На следующем этапе анализа вычисляется определенный вес для каждого элемента матрицы, который и будет отражать степень важности и значимости каждого термина как в документе, в котором он употребляется, так и в наборе документов. Вычисляется вес следующим образом [5]:

$$a_{ij} = L(i, j)\, G(i),$$

где $L(i, j)$ – весовая функция для термина i в документе j и $G(i)$ – весовая функция в целом для термина i. Обычно $L(i, j)$ – бинарная, частотная или логарифмическая функция [5]. Бинарная функция имеет форму:

$$L(i,j) = \begin{cases} 0, \forall \tau_{ij} = 0; \\ 1, \forall \tau_{ij} > 0, \end{cases}$$

где τ_{ij} – частота появления термина i в j-м документе, т. е. в случае *бинарной интерпретации* функция $L(i,j)$ принимает значение 1, если термин i встречается в j-м документе. Если не встречается вовсе, то значение функции равно нулю; а

$$G(i) = \sqrt{\frac{1}{\sum_j \tau_{ij}^2}}.$$

В случае *частотной интерпретации* значение функции $L(i,j)$ – частота τ_{ij} $L(i,j) = \tau_{ij}$; а

$$G(i) = \sqrt{\frac{1}{\sum_j \tau_{ij}^2}}.$$

Логарифмическая интерпретация подразумевает извлечение логарифма из частоты: $L(i, j) = log_2(\tau_{ij}+1)$, а весовая функция приобретает форму:

$$G(i) = \log_2\left(\frac{n}{\gamma_i}\right) + 1,$$

где γ_i – частота появления термина i во всем наборе документов; n – общее число документов.

После того, как матрица векторного пространства создана и взвешена, необходимо ее аппроксимировать. Для аппроксимации ЛСА используют декомпозицию с сингулярными (особенными) значениями (SVD-анализ). Эта методика, близка к факторному анализу.

Следующим этапом проводится масштабирование, при котором матрица «термин-документ» преобразуется в набор k (обычно от 100 до 300) ортогональных коэффициентов (факторов), с помощью которых первоначальная матрица аппроксимируется линейной комбинацией этих факторов [1]. Вместо представ-

ления документов и терминов непосредственно (как векторов независимых слов), ЛСА представляет их как непрерывные значения на каждом из k ортогональных индексирующих измерений, полученных из SVD-анализа.

При формировании корпоративных ИХ будем различать следующие две основные группы технологий быстрого поиска (по территориальному признаку): локальные и корпоративные. Введем также следующие две категории технологий быстрого поиска (по широте предметной области): проблемно-ориентированные и интегральные (т. е. комплексные, по нескольким крупным проблемам).

Имеющийся у авторов данной работы опыт позволяет рекомендовать указанную в таблице предпочтительность использования средств оценки релевантности данных из ИХ для разных групп и категорий технологий быстрого поиска.

ТАБЛИЦА

	Локальные	Корпоративные
Проблемно-ориентированные	ЛСМ	ЛСМ, ЛСА
Интегральные	ЛСМ, ЛСА	ЛСМ, ЛСА, SVD

В данной работе предлагается шкала предпочтительности использования различных комбинаций подобных методов в зависимости от характера анализируемого документа и поставленных пользователем целей. Предлагаемая шкала приоритетов методов определения релевантности актуальна для проблемно-ориентированных технологий ускоренного поиска данных при формировании и эксплуатации корпоративных информационных хранилищ.

1. *Ландэ Д.* Навигация в сети: каталоги, поисковики, порталы // InternetUA. –2000. – № 1 (сентябрь). – С. 43– 47.
2. *Ньюман М.* А теперь займемся посещаемостью // Internet UA. – 2001. – № 11. – С. 48– 52.
3. *Зелинский С.* Виртуальные ищейки // Мир связи, 2001. – № 6. – С. 42– 47.
4. *Herrera F. & Herrera-Viedma E.* Aggregation operators for linguistic weighted information // IEEE Transactions on Systems? Man? And Cybernetics. – 1997. – № 27. – P. 646– 656.
5. *Эйсмонт Ю.* Виртуальная гавань // Мир связи, 2001. – № 10. – С. 62– 64.

Получено 01.07.2003