Система рубрикации данных «Синдбад»

© В.В. Рыбинкин
ООО «Бюро Интернет Технологий БИТ»
ryb@2bit.ru

Иерархическая рубрикация

- Рекурсивный обход дерева рубрик с выполнением на каждом уровне иерархии полнотекстового поиска по терминам, характеризующим ту или иную рубрику.
- Дублирование данных в различных узлах дерева
- Трудность редактирования данных
- Резкий рост объемов данных
- Сложности выбора наиболее адекватного окружения для размещаемого или искомого узла.

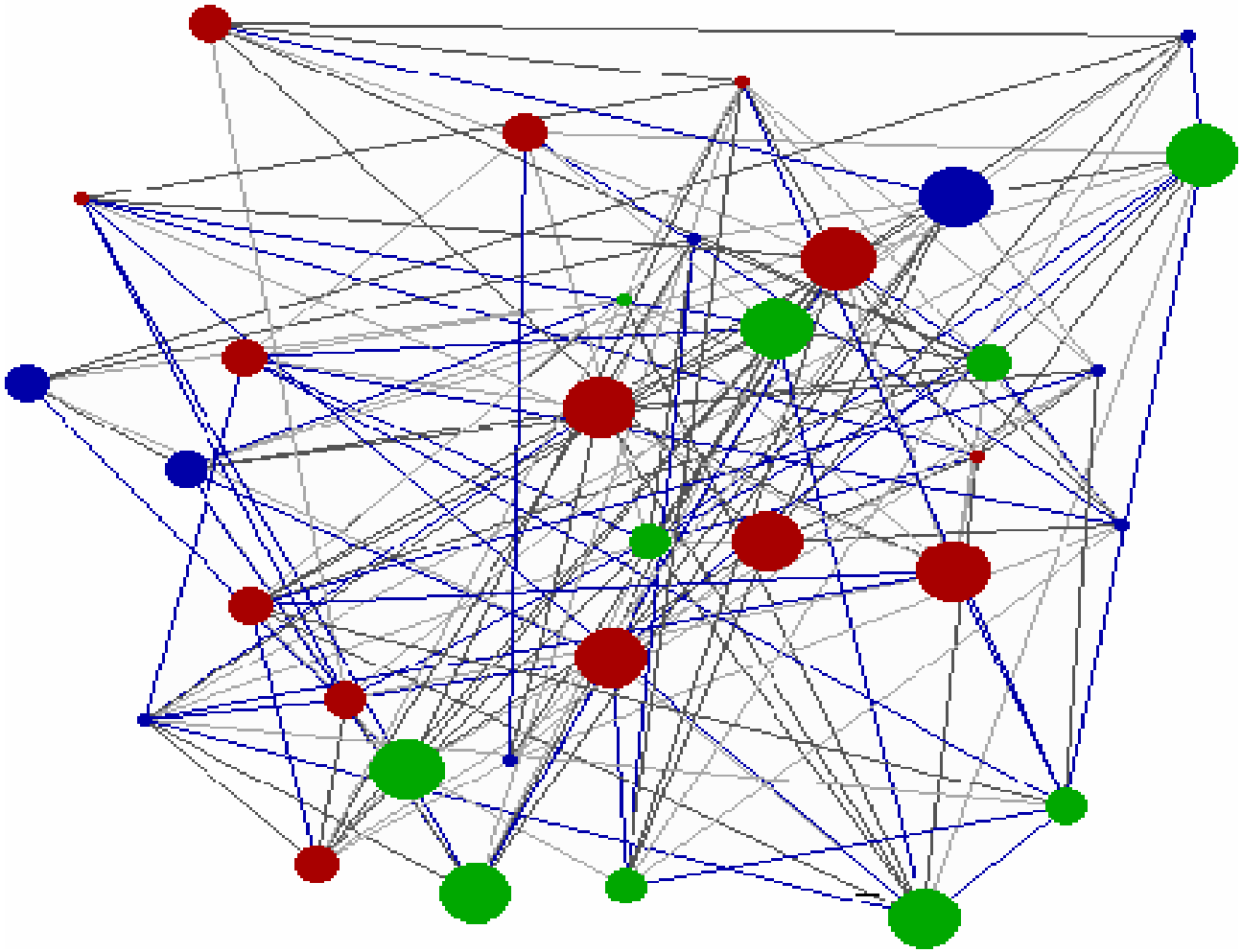
где искать **Автоспорт** - в **Автомобилях** или в **Спорте**?

Основные определения

- **Узел** графа рубрик есть именованная структура данных. Под именем узла понимаем его идентификатор (ID).
- **Ребро** графа есть одно из полей узла типа «ссылка», в которой расположен ID узла, связанного с ним.
- **Тип ребра** графа или **измерение** есть именованная характеристика связи между узлами, представленная этим ребром, например: тематическая, географическая, темпоральная, алфавитная и т.д.
- Количество измерений в узле назовем **размерностью узла**, общее количество измерений во всех узлах графа – **размерностью графа**. Граф называется **многомерным**, если его ребра расположены в разных измерениях, т.е. размерность графа больше единицы.

Представление массива рубрик

Рис. 1. Представление массива рубрик в виде графа.



Алгоритмы рубрикации и реструктуризации данных

Считается, что это алгоритмически дорого и сложно:

- обход дерева – экспоненциальная сложность
- обход графа произвольной структуры, при отсутствии ограничений на количество ребер в узле - с трудом поддается количественной оценке

Разработанный алгоритм

- **линейную** по числу узлов трудоемкость обхода графа
- гарантируется доступ ко всем узлам графа (в частности, если эти связи повреждены)

На практике:

- Трудоемкость линейна **в худшем случае**
- Для некоторых типов задач может оказаться меньше линейной в разы и даже на порядки (в наших экспериментах – на 2-3 порядка)

Алгоритм (2)

- Каждая рубрика характеризуется набором ключевых слов, связанных отношениями булевой логики и скобками для указания приоритета операций.
- Никакого синтаксического, орфографического, семантического анализа рубрицируемых данных **не производится**.
- Анализ словоформ не производится
- Алгоритм ориентирован на производительность
- Усложнение описания рубрики резко улучшает качество, и при достаточно объемном и многокритериальном ее описании дает результаты, близкие к ручному рубрицированию.
- Ограничений по количеству рубрик, к которым может быть приписан элемент, не предусматривается.

Преимущества метода

- высокая производительность, обусловленная простотой алгоритма рубрикации и линейным алгоритмом обхода графа рубрик;
- высокая точность, обусловленная возможностью задания сколь угодно сложной комбинации критериев, характеризующих принадлежность элемента рубрике;
- простота реструктуризации графа данных и повторной их рубрикации по измененному графу;
- не требуется разработка сложных утилит анализа информации, смысловых тезаурусов и т.п.;
- не требуется предварительного обучения и, следовательно, дорогостоящей подготовки обучающих последовательностей экспертами в проблемных областях;

Алгоритм реструктуризации

- Граф рассматривается как таблица неоднородных кортежей.
- Обход графа производится последовательно, элемент за элементом, «не обращая внимания» на связи между ними.
- Реальная трудоемкость обхода оказывается обычно заметно ниже линейной, поскольку имеется возможность просматривать не все, а лишь необходимые элементы (в этом смысле граф рассматривается как отношение со встроенными индексами).

Каталогизация результатов

- Представление данных в виде системы каталогов резко повышает наглядность информации, что особенно важно для пользователей с относительно слабой подготовкой.
- Каталог одновременно является средством эффективного поиска информации без составления специальных запросов к СУБД или средством уточнения таких запросов.
- Каталогизация результатов поиска способна заметно облегчить работу с данными.
- Наглядность представления информации повышает вероятность визуального выявления ошибок в данных.

Выявление ошибок в данных

Задача рубрикации сильно осложняется наличием ошибок в данных. Однако та же рубрикация помогает выявлять эти ошибки.

Если в некоторой коллекции данных 100 раз встречается значение **ААлександр**, 200 раз **Адександр** и 300 раз **Алекандр**. После выделения уникальных значений, каждое из этих слов будет встречаться в справочном массиве лишь однажды, и их нетрудно исправить вручную После обратного преобразования все 600 ошибок будут исправлены.

Дополнительные результаты

- сформирован «сетевой вариант» классификатора компании КОДЕКС в виде графа с сетевой организацией данных (1740 рубрик, http://www.2bit.ru/KODEKC/)
- выявлен ряд грамматических ошибок в наиболее популярных словах документов, представленных для рубрикации (http://www.2bit.ru/KODEKC/errors.zip)

Система рубрикации данных «Синдбад»

© В.В. Рыбинкин
ООО «Бюро Интернет Технологий БИТ»
ryb@2bit.ru