

ПРОБЛЕМЫ ПОИСКА МУЛЬТИМЕДИЙНЫХ ДАННЫХ НА ОСНОВЕ СЕМАНТИЧЕСКИХ СЕТЕЙ

Колб Д. Г.

Кафедра интеллектуальных информационных технологий, Белорусский государственный университет
информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: kolb@bsuir.by

Приводятся основные подходы к организации поиска мультимедийных данных с использованием баз знаний и методов искусственного интеллекта. Перечислены основные проблемы существующих методов поиска мультимедийных данных, в основе которых лежат онтологии. Сформулированы требования к модели поиска, которая, по мнению автора, должна обеспечить решение ряда существующих проблем мультимедийного поиска.

ВВЕДЕНИЕ

Современные поисковые системы развиваются семимильными шагами и с каждым годом удивляют пользователей всё более широкими возможностями. Не смотря на высокие темпы развития теории поиска, многие задачи поиска до сих пор остаются нерешенными или решаемыми с ограничениями. В связи с появлением огромного количества видов мультимедийных данных и форматов их представления в настоящее время наиболее остро стоит проблема их поиск.

Наиболее распространёнными моделями и методами, использующимися для поиска, являются статистические модели и методы, а также модели Data Mining. Однако, всё чаще при их использовании возникает проблема строгого семантического соответствия запрашиваемых мультимедийных данных найденным. Решение этой проблемы традиционными методами поиска не приводит к качественному результату[1]. В последнее время всё чаще и чаще данную проблему пытаются решать на основе лингвистических онтологий, в частности онтологии английского языка WordNet[2] или семантического аннотирования мультимедийных данных.

1. ПОИСК МУЛЬТИМЕДИЙНЫХ ДАННЫХ НА ОСНОВЕ СИСТЕМ МЕТАДААННЫХ И ОНТОЛОГИЙ

Лингвистическая онтология WordNet состоит из 4 сетей для основных знаменательных частей речи английского языка. Базовой словарной единицей в WordNet является не отдельные слова, а неупорядоченные синонимические множества – синсеты, по сути своей являющиеся узлами сети. Каждый синсет содержит список синонимов или синонимичных словосочетаний и указатели, описывающие отношения между ним и другими синсетами. Слова, имеющие несколько значений, включаются в несколько синсетов и могут быть причислены к различным синтаксическим и лексическим классам. В WordNet используются семантические отношения гиперони-

мии, гипонимии, ISA, АКО, меронимии и антонимии. Особую роль в WordNet играет гипонимия: она позволяет организовывать синсеты в виде семантических сетей. Для разных частей речи родовидовые отношения могут иметь дополнительные характеристики и различаться областью охвата. Для удобства использования WordNet человеком каждый синсет снабжен примерами употребления слов в контексте[2].

В мультимедийном поиске WordNet впервые предложил использовать Джордж Миллер (создатель WordNet). В последствии работы Миллера были развиты Тэнсли [3]. Тэнсли использовал WordNet для построения баз изображений, обеспечивая связь понятие (синсет)-изображение. Для повышения эффективности поиска он, используя латентно-семантический анализ (LSA), строил семантический слой, в котором каждая связь синсет-изображение уточнялась весом, указывающим смысловую близость концепта онтологии и изображения. После этого неразмеченные изображения (тестовые изображения) аннотировались на основе сравнения их с обучающей выборкой изображений, используя метод k ближайших соседей. В дальнейшем веса связей понятие-изображение уточнялись на основе анализа отношений между понятиями онтологии WordNet.

К существенным недостаткам данного подхода и других подходов данного класса можно отнести следующее:

- использование онтологий, которые не предназначены для решения задач подобного класса, что выражается в присутствии в онтологии отношений, используемых только в компьютерной лингвистике ;
- наличие проблем при работе с разными языками (в настоящее время в этой области уже получены значимые результаты (RussNet[4] и EuroWordNet[5]);
- зависимость эффективности подхода от объёма онтологии;
- зависимость эффективности подхода от охвата предметных областей онтологией;

- использования поиска без учета смысла всего поискового запроса;
- отсутствие возможности поиска части изображения;
- ориентация на работу с одним типом мультимедийных данных.

Другим подходом к поиску, который можно использовать только для мультимедийных данных размещенных в web-пространстве, является подход на основе семантического или естественно-языкового аннотирования. Основная идея подхода состоит в том, что в процессе создания контента сайтов наряду с мультимедийной информацией присутствует текстовая составляющая, которая описывает то, что представлено в мультимедийных данных. Такая текстовая информация может включать или не включать метаданные, аннотирующие мультимедиа-контент. В подобных наборах данных можно использовать методы поиска текстовой информации на основе статистических моделей или Text Mining, которые обеспечивают более релевантный запросу поиск, чем методы поиска на мультимедийных данных.

К существенным недостаткам данного подхода относятся следующие :

- не все web-документы с мультимедийными данными содержат текст аннотирующий эти данные или текст не находится рядом с изображением;
- наличие текстов разных языков в рамках одного документа, что существенно усложняет задачу релевантного поиска мультимедийных данных;
- отсутствие возможности поиска части изображения.

Наряду с данными подходами, часто используются методы структурно-пространственного анализа web-документов[1], которые позволяют свести задачу поиска мультимедийных в web-документах, содержащих разнородные мультимедийные данные к задаче поиска однородных мультимедийных данных с использованием подходов, описанных выше.

II. ТРЕБОВАНИЯ К МОДЕЛИ ПОИСКА МУЛЬТИМЕДИЙНЫХ ДАННЫХ НА ОСНОВЕ СЕМАНТИЧЕСКИХ СЕТЕЙ

Сформулируем требования, которые на взгляд автора позволят эффективно организовать поиск мультимедийных данных с использованием семантических сетей:

1. наличие системы метаданных для спецификации мультимедийной информации;
2. наличие унифицированного способа кодирования семантической сети, который об-

ладает возможностью гибкого расширения под конкретные прикладные задачи;

3. наличие специализированной онтологии, построенной на основе системы метаданных, в которой используются отношения, позволяющие описать межпредметные связи, связи между мультимедийными данными (их характеристики и свойства) связи между понятиями предметной области, в рамках которой осуществляется поиск;
4. наличие объёмной обучающей базы мультимедийных данных, которая интегрирована в мультимедийную онтологию;
5. наличие методов преобразования мультимедийных данных в различные форматы их представления;
6. наличие языка запросов для вопросно-ответного поиска, в котором обеспечивается комплексная классификация поисковых запросов.

Обеспечение этих требований в конкретных поисковых моделях позволит повысить эффективность поиска мультимедийных данных. По мнению автора, эффективность в первую очередь будет обеспечена за счет использования более специализированных онтологий и языка запросов, который позволит сузить диапазон возможных нерелевантных запросу ответов, за счет комплексной классификации запросов.

1. Magalhaes, J. Statistical models for semantic-multimedia information retrieval: PhD thesis // J. Magalhaes; The Open University. 2008 [Электронный ресурс]. – 2012. Режим доступа: <http://www.doc.ic.ac.uk/~sruieger/www-pub/j.magalhaes-phd.pdf>. – Дата доступа: 27.10.2012
2. WordNet a lexical database for English [Электронный ресурс] / Princeton University . – 2013. – Режим доступа: <http://wordnet.princeton.edu/>. – Дата доступа: 21.09.2013
3. Tansley, R. The multimedia thesaurus: Adding a semantic layer to multimedia information. PhD Thesis // R. Tansley / University of Southampton. – Southampton, UK, 2000. [Электронный ресурс]. – 2013. Режим доступа: <http://eprints.soton.ac.uk/253833/2.hasCoversheetVersion/thesis.pdf>. – Дата доступа: 21.09.2013
4. Проект RussNet [Электронный ресурс] / Кафедра математической лингвистики Филологического факультета Санкт-Петербургского государственного университета . – 2013. – Режим доступа: http://project.phil.spbu.ru/RussNet/index_ru.shtml. – Дата доступа: 21.09.2013
5. EuroWordNet [Электронный ресурс] / University of Amsterdam. – 2013. – Режим доступа: <http://www.illc.uva.nl/EuroWordNet/>. – Дата доступа: 21.09.2013