КОМБИНИРОВАННЫЙ ПОДХОД К ПРЕДСТАВЛЕНИЮ СОДЕРЖАНИЯ И ТЕКСТОВОГО ОПИСАНИЯ МУЛЬТИМЕДИА

Дмитрий Ночевнов

Аннотация: *В статье рассмотрена проблема семантической разницы между содержимым мультимедиа и его текстовым описанием, определяемым вручную. Предложен комбинированный подход к представлению семантики мультимедиа, основанный на объединении близких по содержанию и текстовому описанию мультимедиа в классы, содержащие обобщённые описания объектов, связей между ними и ключевых слов текстовых метаданных из некоторого тезауруса. Для формирования этих классов используются операции иерархической кластеризации и машинного обучения. Данный подход позволяет расширить область поиска и навигации мультимедиа благодаря привлечению медиа-данных, имеющих схожее содержание и текстовое описание.*

Ключевые слова*: multimedia data mining, content based image retrieval, text-based image retrieval.*

ACM Classification Keywords*: H.3.1 Content Analysis and Indexing - Abstracting methods. H.5.1 Multimedia Information Systems - Evaluation/methodology.*

Conference*: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

Интерес к области обработки мультимедиа обусловлен быстром ростом сети WWW и возникающей необходимости быстрой индексации и поиска медиа данных среди большого и разнородного массива информации. Производство и потребление мультимедийного содержания и документов стали обычной практикой благодаря существованию эффективных инструментов создания метаданных в машиночитаемом текстовом формате или же в форме визуальных и аудио дескрипторов (например, в формате MPEG-7), и индексации мультимедиа. Это облегчает их обработку поисковыми машинами и интеллектуальными агентами [Stamou, 2005]. Однако остаётся нерешённой проблема семантического барьера между такими низкоуровневыми характеристиками мультимедиа, такими как цвет, текстура, сцена и т.п., и высокоуровневыми концепциями типа «горный ландшафт», «аномальное поведение человека», используемыми человеком для описания мультимедиа. Преодоление этого барьера является одной из задач Multimedia Data Mining [Stamou, 2005], [Petrushin, 2007].

Традиционный поиск мультимедиа принято разделять на два основных вида [Вихровский, 2006], [Goodrum, 2000] :

1. Поиск на основе текстового описания (Text-Based Retrieval). Данный поиск использует высокоуровневую информацию, опираясь на ключевые слова некоторого тезауруса или свободную текстовую аннотацию.

2. Контентно-зависимый поиск (Content-Based Retrieval), который опирается на использование низкоуровневой информации и визуальных данных в качестве запроса.

Основной проблемой Text-Based Retrieval является сложность точного и полного текстового описания мультимедиа, которое разные пользователи могут идентифицировать по разному исходя из собственного опыта и знаний [Goodrum, 2000]. Одним из решений этой проблемы является Content-Based Retrieval на основе шаблона искомого мультимедиа. Однако в некоторых случаях только на основе низкоуровневого

описания сложно автоматически определить интересующую искателя информацию, например создателя мультимедиа, точное время суток, объекты «за кадром», трёхмерные отношения между изображёнными объектами и т.д. Кроме этого сложно обеспечить точность и полноту такого поиска из-за разнообразия отображения схожих по текстовому описанию мультимедиа в разных предметных областях.

Исходя из этого целесообразно использование комбинированного подхода к индексации и поиску мультимедиа, объединяющего низкоуровневое описание мультимедиа и более высокоуровневое его текстовое описание [Goodrum, 2000]. Это должно способствовать повышению эффективности обработки и поиска мультимедиа благодаря привлечению в процесс поиска сведений о мультимедиа, имеющих схожее содержание и текстовое описание.

1. Комбинированная модель представления знаний мультимедиа

В большинстве своём медиа-данные фактически содержат визуальные или звуковые следы физических объектов, записанные с помощью сенсоров [Stamou, 2005]. Если таких объектов несколько, то сохраняется также информация об их пространственных (в случае изображений и видео данных) или временных (в случае аудио и видео данных) взаимосвязях. Для обозначения этих объектов можно использовать существующие онтологии и тезаурусы, содержащие названия объектов мультимедиа, например Getty's Art and Architecture Thesaurus [AAT], состоящий из более чем 120000 терминов для описания искусства, архитектуры и других культурных объектов, или же Library of Congress Thesaurus of Graphic Materials [LCTGM], и хранить вместе с названием соответствующие метаданные в формате MPEG-7, характеризующие данный объект. Подобный подход к представлению знаний мультимедиа можно найти в работе [Petridis, 2004], в которой для хранения описания MPEG-7 мультимедиа используется онтология в формате Semantic Web. Для формализации связей между объектами можно использовать язык предикатов первого уровня.

Всё это позволит автоматизировать выделение объектов и связей между ними во время индексации мультимедиа, а также расширить область поиска путём привлечения близких по смыслу объектов.

1.1. Модель содержания мультимедиа

Одним из подходов к представлению мультимедиа является формализм семантических сетей [Petridis, 2004], [Dance, 1996]. Следуя ему, для формализации объектов и связей, отображаемых в мультимедиа **m**, и его текстового описания, будем использовать графа, содержащий:

1) множество вершин $v_i \in V_m$, представляющих объекты, отображённые в мультимедиа **m**; каждой из вершин ставится в соответствие некоторое ключевое слово $d(v_i)$ из тезауруса наименований объектов, такое что значения MPEG-7 дескрипторов этих объектов близки к значениям дескрипторов объектов тезауруса;

2) множество связей между двумя вершинами $a_{jk} \in A_m$, соответствующие реальным связям между объектами, отображёнными в мультимедиа; каждой связи ставится в соответствие предикат $p(a_{jk})$;

3) текстовое описание мультимедиа T_m, представляемое в виде множества ключевых слов, и определяемое вручную, или автоматически путём индексации текстового описания мультимедиа.

Одному ключевому слову тезауруса d может соответствовать несколько вершин моделей одной или нескольких мультимедиа в случае повторения одного и того же типа объекта. Один и тот же тип связи, обозначенный предикатом p, также может быть представлен в моделях одной или нескольких мультимедиа.

1.2. Модель класса мультимедиа

Семантически близкие мультимедиа $m_n \in M_c$ предлагаем объединять в классы ***c***, содержащие обобщённую информацию о типичных объектах и связях, отображаемых в мультимедиа, а также ключевые слова их текстовых описаний.

Для его формализации также используем граф, содержащий:

1) множество взвешенных вершин $v_i \in V_c$, описывающих представленные во множестве объединяемых мультимедиа $m_n \in M_c$ объекты; каждой из вершин ставится в соответствие некоторое ключевое слово $d(v_i)$ из тезауруса наименований объектов, а также вес $w(v_i)$, обозначающий количество повторений данного объекта во множестве M_c;

2) множество взвешенных связей между двумя вершинами $a_{jk} \in A_c$, описывающих представленные во множестве объединяемых мультимедиа $m_n \in M_c$ связи между объектами; каждой из связей ставится в соответствие некоторый предикат $p(a_{jk})$, а также вес $w(a_{jk})$, обозначающий количество повторений данной связи между объектами во множестве M_c;

3) текстовое описание класса мультимедиа T_c, определяемое индексацией текстовых описаний t_m множества мультимедиа $m_n \in M_c$; представляет собой множество взвешенных ключевых слов из текстовых описаний t_m с весами $w(t_i)$, обозначающими количество повторений данного ключевого слова среди t_m.

2. Последовательность обработки мультимедиа

Можно предложить следующий алгоритм обработки мультимедиа:

1. Индексация нового мультимедиа и составление модели содержания мультимедиа m:

а) автоматическое выделение объектов, отражённых в мультимедиа и поиск наиболее близких по значениям MPEG-7 дескрипторов ключевых слов d тезауруса объектов;

б) автоматическое выделение связей между объектами и поиск соответствующих этим связям предикатов p, и составление модели содержания мультимедиа;

в) определение текстового описания t_m.

2. Определение семантически наиболее близкого к мультимедиа класса c_i и обновление его модели данными о новом мультимедиа, или же добавление нового класса мультимедиа.

3. Проверка качества разбиения мультимедиа на классы и при необходимости кластеризация семантически однородных классов.

4. Использование во время поиска сведений о близких по содержанию и описанию мультимедиа из модели класса мультимедиа.

Семантически наиболее близкий класс c_i или множество классов C могут быть определены пользователем при навигации в базе знаний мультимедиа, или же вычислены автоматически путём анализа расстояния между мультимедиа m и существующими классами по формуле:

$$c = c_k, \text{при } h(m, c_k) = \min(h(m, c_k)) \le h_{\max}, i = \overline{1, N_c}, \quad (1)$$

где N_c – общее кол-во классов мультимедиа.

3. Формализация расстояния между классами и экземплярами мультимедиа

Определим метод вычисления расстояния между классами и экземплярами мультимедиа. Для предложенных в предыдущем разделе моделей наиболее подходят модифицированные способы

вычисления сходства, основанные на описании характеристик объекта в виде вектора ключевых слов и вычислении совпадения элементов векторов и их весов [Озкархан, 1989].

Согласно определению из [Дюран, 1977]:

Определение 1. Неотрицательная вещественная функция $z(x, y)$ называется *функцией близости*, если:

1) $0 \le z(x, y) < 1$ для $x \ne y$,

2) $z(x, x) = 1$,

3) $z(x, y) = z(y, x)$.

Определение 2. Неотрицательная вещественная функция $h(y, x) = 1 - z(y, x)$ называется *функцией расстояния*.

При вычислении семантически близкого класса мультимедиа необходимо определять расстояние между классом и экземпляром мультимедиа $h(m,c)$. Для его расчёта будем учитывать совпадение объектов моделей, связей между ними, ключевых слов в текстовых описаниях и их весов:

$$h(m,c) = 1 - z(m,c) = 1 - \frac{\lambda_v \cdot z_v(m,c) + \lambda_a \cdot z_a(m,c) + \lambda_t \cdot z_t(m,c)}{\lambda_v + \lambda_a + \lambda_t}, \qquad (2)$$

где $\lambda_v, \lambda_a, \lambda_t \in [0,1]$- коэффициенты влияния, определяемые опытным путём,

$$z_v(m,c) = \left(1 - \frac{\sum_{\forall v \in (V_m \cap V_c)} n[w_c(v)]}{card(V_m \cap V_c)}\right) \cdot \frac{card(V_m \cap V_c)}{card(V_m \cup V_c)}$$ – степень близости по составу объектов,

$$z_a(m,c) = \left(1 - \frac{\sum_{\forall a \in (A_m \cap A_c)} n[w_c(a)]}{card(A_m \cap A_c)}\right) \cdot \frac{card(A_m \cap A_c)}{card(A_m \cup A_c)}$$ – степень близости по составу связей между объектами,

$$z_t(m,c) = \left(1 - \frac{\sum_{\forall t \in (T_m \cap T_c)} n[w_c(t)]}{card(T_m \cap T_c)}\right) \cdot \frac{card(T_m \cap T_c)}{card(T_m \cup T_c)}$$ – степень близости текстовых описаний мультимедиа,

$n(x) = \left|0.1 \cdot \lg(1+x)\right|_{\mathrm{mod}1}$- функция нормализации весов модели класса мультимедиа,

$\lambda_v = 0$, если $card(V_m \cup V_c) = 0$; $\lambda_a = 0$, если $card(A_m \cup A_c) = 0$; $\lambda_t = 0$, если $card(T_m \cup T_c) = 0$.

При объединении множества близких по содержанию мультимедиа в класс возникает потребность в определении расстояния между отдельными экземплярами мультимедиа $h(m_k,m_l)$. Для его расчёта будем учитывать совпадение объектов и связей, отображаемых в мультимедиа m_k и m_l, а также совпадение ключевых слов их текстовых описаний:

$$h(m_k,m_l) = 1 - z(m_k,m_l) = 1 - \frac{\lambda_v \cdot z_v(m_k,m_l) + \lambda_a \cdot z_a(m_k,m_l) + \lambda_t \cdot z_t(m_k,m_l)}{\lambda_v + \lambda_a + \lambda_t}, \qquad (3)$$

где $z_v(m_k,m_l) = \frac{card(V_{m_k} \cap V_{m_l})}{card(V_{m_k} \cap V_{m_l})}$ – степень близости по составу объектов,

$z_a(m_k,m_l)=\dfrac{card(A_{m_k}\cap A_{m_l})}{card(A_{m_k}\cap A_{m_l})}$ – степень близости по составу связей между объектами,

$z_t(m_k,m_l)=\dfrac{card(T_{m_k}\cap T_{m_l})}{card(T_{m_k}\cap T_{m_l})}$ – степень близости текстовых описаний мультимедиа,

$\lambda_v=0$, если $card(V_{m_k}\cup V_{m_l})=0$; $\lambda_a=0$, если $card(A_{m_k}\cup A_{m_l})=0$; $\lambda_t=0$, если $card(T_{m_k}\cup T_{m_l})=0$.

Для определения качества разбиения множества мультимедиа на классы следует вычислять расстояние между классами мультимедиа $h(c_i,c_j)$. Будем определять его с учётом совпадение объектов классов, связей между ними, ключевых слов в текстовых описаниях и их весов:

$$h(c_i,c_j)=1-z(c_i,c_j)=1-\frac{\lambda_v\cdot z_v(c_i,c_j)+\lambda_a\cdot z_a(c_i,c_j)+\lambda_t\cdot z_t(c_i,c_j)}{\lambda_v+\lambda_a+\lambda_t}, \quad (4)$$

где $z_v(c_i,c_j)=\left(1-\dfrac{\sum\limits_{\forall v\in(V_{c_i}\cap V_{c_j})} n(\left|w_{c_i}(v)-w_{c_j}(v)\right|)}{card(V_{c_i}\cap V_{c_j})}\right)\cdot\dfrac{card(V_{c_i}\cap V_{c_j})}{card(V_{c_i}\cup V_{c_j})}$ – степень близости по составу объектов,

$z_a(c_i,c_j)=\left(1-\dfrac{\sum\limits_{\forall a\in(A_{c_i}\cap A_{c_j})} n(\left|w_{c_i}(a)-w_{c_j}(a)\right|)}{card(A_{c_i}\cap A_{c_j})}\right)\cdot\dfrac{card(A_{c_i}\cap A_{c_j})}{card(A_{c_i}\cup A_{c_j})}$ – степень близости по составу связей,

$z_t(c_i,c_j)=\left(1-\dfrac{\sum\limits_{\forall t\in(T_{c_i}\cap T_{c_j})} n(\left|w_{c_i}(t)-w_{c_j}(t)\right|)}{card(T_{c_i}\cap T_{c_j})}\right)\cdot\dfrac{card(T_{c_i}\cap T_{c_j})}{card(T_{c_i}\cup T_{c_j})}$ – степень близости текстовых описаний,

$\lambda_v=0$, если $card(V_{c_i}\cup V_{c_j})=0$; $\lambda_a=0$, если $card(A_{c_i}\cup A_{c_j})=0$; $\lambda_t=0$, если $card(T_{c_i}\cup T_{c_j})=0$.

4. Обновление класса мультимедиа

Обновление класса мультимедиа сведениями о новом мультимедиа предлагаем делать в режиме обучения, увеличивая веса только тех объектов, связей между ними и ключевых слов текстового описания, которые повторяются в новом мультимедиа. В соответствии с этим правилом во время обучения класса ***c*** информацией о мультимедиа ***m*** последовательно выполняется:

1) модификация значений весов вершин модели класса $w_c(v_i)$ по формуле:

$$\forall v_i\in V'=V_m\cap V_c\rightarrow w_c(v_i)=w_c(v_i)+1;$$

2) модификация значений весов связей объектов $w_c(a_i)$ по формуле:

$$\forall a_i\in A'=A_m\cap A_c\rightarrow w_c(a_i)=w_c(a_i)+1;$$

3) модификация значений весов ключевых слов текстового описания класса $w_c(t_i)$ по формуле:

$$\forall t_i\in T'=T_m\cap T_c\rightarrow w_c(t_i)=w_c(t_i)+1;$$

4) дополнение класса ***c*** недостающими объектами, связями и ключевыми словами текстового описания с весами, равными 1:

$$c = c + (m \setminus c) = \{V_c = V_c + V_m \setminus V_c, A_c = A_c + A_m \setminus A_c, T_c = T_c + T_m \setminus T_c\}.$$

В случае, если не будет найден семантически близкий класс, информация о новом мультимедиа может быть добавлена в базу знаний в виде нового класса с единичными значениями весов объектов, связей и ключевых слов *t*.

5. Автоматическая классификация мультимедиа

Для автоматической классификации мультимедиа и разбиения их на классы предлагаем использовать *собирающий метод иерархической кластеризации* [Pedrycz, 2005]. В соответствии с ним кластеризация будет начинаться с единственных кластеров для каждого мультимедиа, которые затем объединяются в кластеры, формируя двухуровневую иерархическую структура, в которой:

1-ый уровень– кластеры C_i с минимальным расстоянием между кластерами $h_{\min}=1$;

2-ой уровень – подкластеры C_{ij} с минимальным расстоянием между кластерами $h_{\min}<1$.

Разбиение мультимедиа на подкластеры второго уровня предлагаем выполнять пошагово с изменением значения h_{min} от 0 до 1, пока не будет достигнуто приемлемое качество разбиения. Для его оценки предлагаем использовать меру внутренней однородности кластера η_o и меру разнородности кластеров η_m, вычисляемые следующим образом:

$$\eta_{o_i} = \begin{cases} \dfrac{2}{N_i(N_i-1)} \sum_{j=1}^{N_i} \sum_{k=j+1}^{N_i} h(m_j, m_k), \text{если } N_i > 1 \\ 1, \text{если } N_i = 1 \end{cases}$$, где N_i – кол-во мультимедиа, объединённых в кластер C_i,

$$\eta_m = \begin{cases} \dfrac{2}{N_C(N_C-1)} \sum_{j=1}^{N_C} \sum_{k=j+1}^{N_C} h(m_ц^{\langle C_j \rangle}, m_ц^{\langle C_k \rangle}), \text{если } N_C > 1 \\ 0, \text{если } N_C = 1 \end{cases},$$

где $m_ц^{\langle C_i \rangle}$ – центральное мультимедиа кластера C_i, для которого выполняется условие $\sum_j h(m_j, m_ц) = \min, \forall m_j \in C_i$,

N_C – кол-во анализируемых кластеров.

По результатам кластеризации выполняется группировка мультимедиа $m_i \in C_j$ в класс c_i. При группировке можно использовать операцию обновления класса сведениями о мультимедиа, описанную в пункте 4 данной статьи.

Выводы

Хотя на сегодняшний день производство и потребление мультимедийного содержания и документов стали обычной практикой, остаётся нерешённой проблема семантической разницы между содержимым мультимедиа и его текстовым описанием, определяемым пользователями. В статье обоснована целесообразность использования комбинированного подхода к индексации и поиску мультимедиа, объединяющего низкоуровневое описание мультимедиа и более высокоуровневое его текстовое описание. Предложен комбинированный подход к представлению семантики мультимедиа, основанный на объединении близких по содержанию и текстовому описанию мультимедиа в классы, содержащие

обобщённые описания объектов, связей между ними и ключевых слов текстовых метаданных из некоторого тезауруса. Для формирования этих классов используются операции иерархической кластеризации и машинного обучения.

Описанный подход даёт возможность расширить область поиска и навигации мультимедиа благодаря привлечению медиа-данных, имеющих схожее содержание и текстовое описание.

Литература

[AAT] Getty's Art and Architecture Thesaurus. - http://www.getty.edu/research/conducting_research/vocabularies/aat/

[Dance, 1996] Dance Sandy, Caelli Terry, Liu Zhi-Qiang. Picture Interpretation: A Symbolic Approach//World Scientific Series In Machine Perception And Artificial Intelligence; Vol. 20 – 1996.

[Goodrum, 2000] Goodrum A. A. Image information retrieval: An overview of current research // Informing Science, 3(2):P.63-66, February 2000.

[LCTGM] Thesaurus for Graphic Materials // Library of Congress. - http://www.loc.gov/rr/print/tgm1/

[MPEG-7] MPEG-7 Overview. http://www.mpeg-7.com (Industry Focus Group)

[Pedrycz, 2005] Pedrycz Witold. Knowledge-based clustering: From Data to Information Granules. - John Wiley & Sons, 2005. - 316p.

[Petridis, 2004] Petridis K., Kompatsiaris I., Strintzis M.G., Bloehdorn S., Handschuh S., Staab S., Simou N., Tzouvaras V., Avrithis Y. Knowledge representation for semantic multimedia. Content analysis and reasoning. – EWIMT, 2004.

[Petrushin, 2007] Petrushin Valery A. and Khan Latifur (Eds). Multimedia Data Mining and Knowledge Discovery. - Springer-Verlag London Limited, 2007. – 521p.

[Stamou, 2005] Stamou Giorgos and Kollias Stefanos (Eds). Multimedia Content and the Semantic Web. - John Wiley & Sons Ltd, 2005 - 392 p..

[Вихровский, 2006] Вихровский Кирилл, Игнатенко Алексей. Применение MPEG-7 для классификации и поиска визуальных данных // Сетевой журнал "Графика и Мультимедиа". - 23.12.2006 http://cgm.graphicon.ru/content/view/161/61/

[Дюран, 1977] Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977. – 128 с.

[Озкархан, 1989] Озкархан Э. Машины баз данных и управление базами данных: Пер. с англ. – М.: Мир, 1989. – 696с.

Authors' Information

Дмитрий Ночевнов *– доцент кафедры информационных технологий проектирования Черкасского государственного технологического университета.*

Адрес: Кафедра информационных технологий проектирования, Черкасский государственный технологический университет, 18006, г.Черкассы, бул.Шевченка, 460; e-mail: dmitry.ndp@gmail.com