ОЦЕНКА АДЕКВАТНОСТИ МЕТОДОВ ИНТЕГРИРОВАНИЯ СХЕМ ДАННЫХ

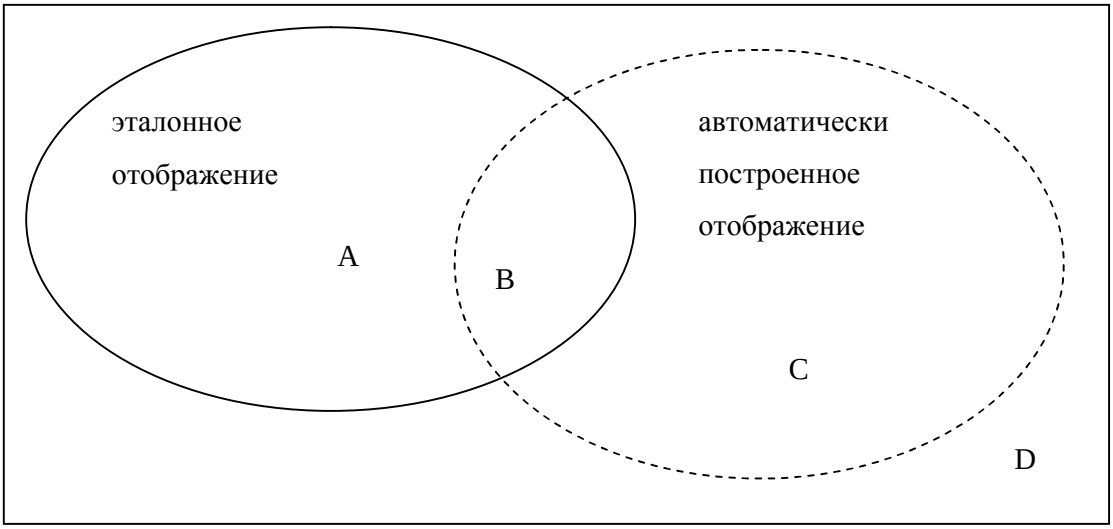
Комар Ф. В.

Липецкий Государственный Технический Университет

Липецк, Россия

В настоящее время ведется активная работа по разработке методов интегрирования схем данных [1, 2]. Предлагаются новые подходы и методы автоматизированного интегрирования [3]. Как следствие встает проблема оценки адекватности методов интегрирования схем данных [4]. Важную роль в оценке результатов применения методов интегрирования схем данных играет эталонное отображение элементов, построенное экспертами.

На базе такого эталонного отображения можно различными методами вычислять количественные оценки качества отображения построенного автоматизированным методом.



A: неверно не отождествленные B: верно отождествленные

C: неверно отождествленные D: верно не отождествленные

Рис. 1. Сравнение эталонного и автоматически построенного отображения элементов схем данных

На рисунке 1 представлены возможные варианты отождествлений. Множество A – это множество истинных, определенных экспертом, соответствий между элементами схем данных. По своей сути множество A – это ошибочно не распознанные соответствия. Множество B – это множество истинных соответствий, которые были включены в автоматически построенное отображение элементов схем данных. По своей сути множество B – это та часть соответствий, которая была верно распознана методом. Множество C – это

множество соответствий, которые были включены в автоматически построенное отображение, но на самом деле не являющихся истинными. По своей сути множество C – это множество ошибочно распознанных соответствий. Множество D – это множество ложных соответствий. По своей сути множество D является множеством верно отброшенных методом соответствий. Очевидно, что чем точнее совпадают множества соответствий эталонного отображения и автоматически построенного отображения, тем выше адекватность автоматически построенного отображения.

Наиболее простыми оценками адекватности построенного отображения могут служить следующие численные характеристики [4]:

$$P = \frac{|B|}{|B| + |C|} \tag{1}$$

Данная оценка отображает долю найденных истинных соответствий по отношению к общему числу соответствий вошедших в автоматически построенное отображение.

$$R = \frac{|B|}{|B| + |A|} \tag{2}$$

Данная оценка отображает долю автоматически найденных истинных соответствий по отношению к общему числу истинных соответствий.

В случае, когда автоматически построенное отображение дает идеальный результат, имеем $P = R = 1$. Однако рассмотренные отдельно друг от друга ни оценка P, ни оценка Q не дают возможности сделать выводы о качестве рассматриваемого отображения. Действительно, оценка P может быть увеличена путем включения в отображения малого числа соответствий с предельно высоким уровнем достоверности. При этом оценка Q будет заведомо занижена. Включая в отображение как можно большее количество соответствий, будет увеличена оценка Q. При этом оценка P будет заведомо снижена.

Представленные ниже оценки лишены указанных выше недостатков:

$$F - Measure(a) = \frac{|B|}{(1-a) * |A| + |B| + a * |C|} = \frac{P * R}{(1-a) * P + a * R} \tag{3}$$

Комбинированная оценка, которая с помощью параметра a позволяет изменять вклад оценок P и Q в конечный результат. При $a = 1$ $F - Measure(a) = P$, оценка Q не учитывает. При $a = 0$ $F - Measure(a) = Q$, оценка P не учитывает. При $a = 0.5$ оценки P и Q учитываются равноценно и можно получить следующую комбинированную оценку [**Ошибка! Источник ссылки не найден.**]:

$$F - Measure = \frac{2 * |B|}{(|A| + |B|) + (|B| + |C|)} = 2 * \frac{P * R}{P + R} \quad (4)$$

В работе [5] была представлена, а в работе [3] использована следующая оценка:

$$Overall = 1 - \frac{|A| + |C|}{|A| + |B|} = \frac{|B| - |C|}{|A| + |B|} = R * \left(2 - \frac{1}{P}\right) \quad (5)$$

С учетом вышесказанного предлагается следующий метод оценки адекватности метода интегрирования схем данных:

Сравнительный анализ методов должен проводиться на заранее подготовленных тестовых задачах. Тестовые задачи должны быть стандартизированы и общепризнанны.

Эталонное отображение в каждой тестовой задаче должно быть построено с учетом мнения нескольких экспертов. Возможно усреднение отображений построенных каждым экспертом для получения единого эталонного отображения. Или же возможно сравнение результатов автоматизированного отождествления с эталонным отображением каждого эксперта и усреднение полученных величин качества.

Для численного анализа результатов предлагается использовать количественные оценки качества $F - Measure$ и $Overall$.

СПИСОК ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ

1. Брюхов Д.О. Интероперабельные информационные системы: архитектуры и технологии. / Задорожный В.И., Калиниченко Л.А., Курошев М.Ю., Шумилов С.С. // СУБД, № 4, 1995
2. Калиниченко Л. А. Методы и средства интеграции неоднородных баз данных. - М.: Наука, 1983. - 423 с.
3. Do Hong-Hai, Rahm Erhard. COMA – A System for Flexible Combination of Schema Matching Approach. // VLDB, 2002.
4. Do Hong-Hai, Melnik Sergey, Rahm, Erhard. Comparison of Schema Matching Evaluations // Proc. GI-Workshop "Web and Databases", Erfurt, Oct. 2002.
5. Melnik Sergey, Garcia-Molina Hector, Rahm Erhard. Similarity Flooding: A Versatile Graph Matching Algorithm (Extended Technical Report) 2001.