УДК 681.3: 612.84

А.А. Стенин, Ю.А. Тимошин, М.М. Ткач, Т.Г. Шемсединов

О МЕТРИКАХ ДЛЯ БАЗЫ ЗНАНИЙ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА

Аннотация: Рассматриваются вопросы построения метрик, которые можно использовать в задачах анализа лингвистических данных при их компьютерной обработке - для построения новых параметрических критериев оценки характеристик обрабатываемого текстового материала и непосредственно самих процессов лингвистического анализа и обработки с использованием технологий баз знаний. БЗ используются как инструмент интеграции промежуточных и окончательных результатов, накопления дополнительной информации и знаний о них, их структуризации и классификации, обнаружения новых фактов и получения новых знаний на их основе.

Ключевые слова: метрики, база знаний, метаданные, лингвистический анализ, обработка текстов, метамодель

Введение

Метрики можно условно разделить на два класса – значимые или информативные и измеримые [1], которые существенно отличаются друг от друга тем, что они позволяют посмотреть на один и тот же процесс под разными углами зрения, поэтому часто используются в комплексе и только так могут служить отправной точкой для принятия объективных решений. Иногда можно расширить класс измеримых метрик и определить для них дополнительно, что они могут быть прямого действия, т.е. непосредственно измеряться при изменении состояния бизнес-процесса или состояния выполняемой задачи в процессе обработки и косвенного оценивания, т.е. вычисляться на основе других показателей процесса, которые можно фактически измерить [1-2].

Для обоих классов метрик должны выполняться следующие условия и ограничения:

1. метрики должны учитывать цели и задачи выполняемого проекта или процесса;
2. Все метрики должны непосредственно следовать из конкретной цели;
3. метрики должны соответствовать принципам и требованиям любого из стандартов -ISO 20000 или ITIL (IT Infrastructure Library) или COBIT или SMART;
4. должна быть разработана и описана адекватная процедура измерения метрики, которая определяет однозначно последовательность действий и точек контроля для получения достоверного конечного результата;
5. метрики могут быть простыми, составными и интегрированными;

 ISSN 1560-8956

6. интегрированные метрики разных процессов должны быть одинаково структурированы для возможности их сопоставления для целей управления и иметь взвешенную оценку для совокупности метрик конкретного процесса;
7. метрика должна входить в определенную модель данных, в которой она используется;
8. метрике могут назначать следующие параметры:
   - Целевое значение
   - Среднее значение
   - Опасное значение
   - Диапазон значений;
9. метрике может быть назначен приоритет (по сравнению с др. метриками процесса), который может со временем меняться;
10. для каждой метрики может быть определен KPI (ключевой показатель эффективности), его максимальное и минимальное значения. Он выражается в относительных величинах, обычно в процентах;
11. для каждой метрики, по возможности, определяется уровень зрелости взаимодействия процессов [2], т.к. со временем наступает момент оптимальности их количества, т.е. когда число низовых метрик минимально, а их замещают составные метрики.

На основе информативных метрик могут быть построены контентные метрики для задач лингвистического анализа.

Метрики создаются для процессов и характеристик [3-4], которые могут меняться в процессе обработки или для которых возникает необходимость их оценивания (как правило, путем прямого или косвенного измерения). Степень такого изменения может служить дополнительным критерием эффективности лингвистической обработки текстов и речи.

База знаний (БЗ) для лингвистического анализа описана в [3] и представляет собой многоуровневую структуру для хранения информации о данных, тематических знаниях, интеллектуальных знаниях и опыте применения с использованием соответствующих метамоделей. Она включает информацию и знания по задачам наполнения БЗ информационными объектами, знаниями о технологиях и схемах их хранения, анализа и обработки, а также по системе управления самой базой знаний.

Поскольку многие исходные ресурсы информационного пространства в задачах лингвистического анализа и обработки текстов и речи, а также их характеристики и оценки для базы знаний формируются децентрализовано, для обеспечения их эффективного использования в систему можно встроить ряд сервисов, позволяющих измерять и визуализировать различные показатели развития и использования информационных ресурсов. Такие сервисы позволят:

автоматически измерять на основе поддерживаемых метаданных и данных, регистрируемых в процессе динамического функционирования БЗ, различные полезные показатели и характеристики процессов анализа или обработки [5];

измерять некоторые показатели метаданных и знаний, описывающих публикуемые и хранимые в системе информационные ресурсы [6-7];

формировать и предоставлять пользователям для анализа указанные показатели, в том числе, с целью оптимизации процессов анализа и обработки, а также с целью выявления новых фактов и формирования новых знаний на их основе.

Метрики могут создаваться для оценки затрат разнородных ресурсов процессов анализа и обработки различного вида лингвистических объектов или фрагментов текста, а также для оптимизации указанных процессов. Эти метрики должны быть наблюдаемыми, т.е. или измеряемыми непосредственно в указанных процессах или вычисляемыми по окончанию анализа или обработки [8-9].

Постановка задачи

Целью данной работы является разработка метрик для измерения или оценивания параметров и характеристик компонентов или процессов лингвистического анализа при динамической обработке, а также создание обобщенной метамодели этих метрик на их основе для компактного представления в БЗ лингвистического анализа.

Выделим основные принципы, которые должны быть заложены в разработку форм представления, описания и формирования наполнения БЗ информационными объектами, знаниями о технологиях и схемах их хранения, анализа и обработки, а также по системе управления самой базой знаний.

Для задач лингвистического анализа и обработки основными будем считать задачи морфологического разбора слов и речи, задачи синтаксического анализа предложения и семантического анализа текста, а также тематические знания по их моделям, опыту применения и использования.

БЗ может использоваться как инструмент интеграции промежуточных и окончательных результатов, накопления дополнительной информации и знаний о них, их структуризации и классификации, обнаружения новых фактов и получения новых знаний на их основе.

Кроме данных, информации, тематических знаний, интеллектуальных знаний по их использованию в нашей БЗ должны храниться и обрабатываться также объекты/субъекты/процессы/явления/события/ситуации. Они выявляются при обработке исходных текстов в виде лингвистических объектов и имеют различного вида связи и отношения между собой, в том числе синтаксические, семантические и ассоциативные.

 ISSN 1560-8956

Обычно большое разнообразие типов объектов и связей, которые выявляются в результате лингвистического анализа в полипредикатных и монопредикатных структурах текстов, порождают необходимость осуществлять обработку огромного числа возможных вариантов структурных связей и получаемых структур этих текстов. В результате лингвисты сталкиваются с задачами перебора вариантов решений большой размерности.

Использование метамоделей для описания, хранения и обработки данных, информации и знаний позволяет создать несколько слоев абстракции и значительно сократить объем перерабатываемой информации и данных в задачах лингвистической обработки [3] текстов с монопредикатной структурой.

В статье мы будем придерживаться предположения, что информация, тематические знания, интеллектуальные знания, а также описание опыта или навыков реализации интеллектуальных знаний [3], которые есть в БЗ – представляют различные *функциональные слои БЗ* для описания, обработки и хранения своих объектов и процессов, а также знаний о них.

Кроме указанных объектов и компонентов, БЗ включает также различные средства хранения (базы данных, репозитории, каталоги, тематические справочники и классификаторы), средства обработки и анализа/выявления новых фактов и знаний, др. системы, которые образуют систему поддержки БЗ.

Метрики функциональных слоев БЗ лингвистических объектов

Для слоя данных метриками могут служить общеизвестные параметры и характеристики – количество или длина исходных данных, результатов, файлов, слов/предложений/абзацев/страниц в тексте, символов или знаков в них, длина в байтах для аудио-видео ряда или изображения, объем файлов и число их типов для multi-media данных, число объектов лингвистического анализа или обработки, т.п. Т.е. они создаются для этого слоя относительно просто. Для других слоев процесс конструирования метрик не такой явный.

Для слоя информации прямыми измеряемыми метриками могут служить следующие характеристики - параметры обработки, параметры источников, время анализа/обработки, контрольная сумма, значение ценности информации. Наблюдаемыми метриками могут служить следущие:

- Число типов представлений в БЗ для каждого вида лингвистических объектов;
- Количество параметров обработки для конкретного типа данных;
- Число метаданных для описания лингвистического объекта определенного вида;
- Число интерфейсов, требуемых для обработки лингвистического объекта;

- Среднее время распаковки фиксированного объема данных;
- Количество словарей и справочников, которые требуется в процессе лингвистического анализа или обработки конкретного лингвистического объекта;
- Число внешних источников данных, используемых для обработки;
- Типы драйверов для взаимодействия с внешними источниками данных;
- Число форматов требуемых в процессе распаковки контента;
- Число характеристик для описания, хранения или обработки конкретного лингвистического объекта;
- Число шаблонов представлений в БЗ для каждого вида лингвистических объектов;
- Количество метаданных описания или хранения структуры данных;
- Число идентификаторов, которые используются в шаблоне представления;
- Число идентификаторов, которые требуются для однозначного описания или обработки конкретного лингвистического объекта;
- Число форматов представления и отображения результатов анализа или обработки;
- Количество ссылок для конкретного лингвистического объекта, которые использованы в процессе анализа или обработки;
- Число индексированных данных в БЗ по типам;
- Число сервисов использованных в процессе анализа или обработки для каждого вида лингвистических объектов (max / min).

Для слоев тематического знания и интеллектуальных знаний можно предложить следующие прямые и вычисляемые метрики:

- Количество актуальных понятий в БЗ (т.е. которые активно используются);
- Число наборов представлений для каждого вида лингвистических объектов;
- Количество представлений в конкретном наборе;
- Количество параметров метамодели;
- Число (типов) метамоделей в БЗ (т.е. описания, хранения, управления);
- Набор параметров алгоритма (число условных и безусловных переходов, число операторов, количество используемых процедур, max число вложений для процедуры);

 ISSN 1560-8956

- Число параметров управления в модели;
- Число моделей различных типов в БЗ для анализа (лингвистического или др. видов);
- Количество технологий разного назначения, описанных в БЗ и системах ее поддержки;
- Число процессов в конкретной задаче (анализа или обработки);
- Число параметров или метрик в критерии (“мерность” критерия);
- Оценка точности реального критерия;
- Характеристики и параметры категорий, видов, классов, которые позволяют их оценивание;
- Параметры требуемых ресурсов (памяти дисковой или процессорной) для хранения и установки классификатора;
- Число шагов обработки для соответствующего правила;
- Число подсистем и ресурсы для каждого этапа обработки, численные характеристики каждого промежуточного результата;
- Сложность схемы обработки - число подсистем и требуемых ресурсов для реализации схемы обработки (число серверов, БД, внешних источников, ресурсы для архивации, характеристики среды обработки, т.п.), количество этапов для реальной схемы обработки;
- Число шаблонов обработки или анализа в БЗ;
- Характеристики программных средств (параметры требуемых ресурсов для загрузки и исполнения [кбайт]);
- Число БД, число таблиц и хранимых процедур в каждой;
- Характеристики среды исполнения – количество интерфейсов, компонентов, протоколов, параметров управления, предельные размеры дискового пространства, параметры ОС и виртуализации;
- Число связей конкретного типа и вида в простом предложении, словосочетании или метамодели; количество ссылок и отношений в сложном предложении/на странице/в тексте;
- Число форм представления и визуализации информации в БЗ (для разных категорий пользователей), частота их использования и число обращений к ним/ число вызовов;
- Число типовых задач анализа или обработки, которые были решены в БЗ, число использованных средств обработки (абсолютное и относительное) и частота использования их за период;
- Число различных актуальных моделей, описывающих результаты анализа или обработки, в БЗ;
- Показатели и оценки ценности информации, если такие разработаны предварительно (например, в случае потери результатов анализа надо повторить обработку, а это прямые дополнительные затраты ресурсов, т.п.);

- Число использованных типов классификаторов для классификации лингвистического объекта;
- Мах число типов представлений в задаче создания набора представлений (для критерия полноты описания лингвистического объекта);
- Эффективность шаблона - процент запросов, которые реализованы с использованием соответствующего шаблона, к общему числу запросов к БЗ;
- Минимальное затраченное число этапов (шагов) и объем ресурсов на обработку для конкретного алгоритма;
- Минимальное число блоков в схеме обработки, при условии полной ее функциональности для конкретной задачи;
- Заданные диапазоны точности для параметрических критериев выполнения процессов анализа и обработки;
- Максимальная близость оценок экспертов между собой для задач выработки и принятия решений;
- Сравнение ряда параметрических оценок при оценивании близости реального и желаемого результатов;
- Число сформированных и записанных новых фактов и знаний в БЗ.

Для слоя *опыта решения задач*, использование функциональности которого позволяет реализовать интеллект, мы не будем рассматривать вопрос построения метрик из-за ограниченности объема статьи.

Некоторые из предложенных метрик могут быть использованы в виде метаданных, которые используются для передачи параметров управления или описания между интерфейсами различных IT - систем динамической обработки текстов.

На основе приведенных метрик можно разрабатывать новые критерии оптимизации числа вариантов для получаемых структур лингвистических объектов при анализе или обработке исходных текстов, а также существенно уменьшить количество этих структур, а значит, и минимизировать требуемые ресурсы для хранения полученных результатов и новых знаний.

Наборы метаданных могут служить основой для формирования метамоделей, которые представляют собой, например, знания о структуре объекта (синтаксическая или морфологическая структуры), найденных связях или отношениях между объектами, признаках или характеристиках, которые можно использовать для классификации их в БЗ, и т.п.

Отдельные метрики можно использовать как метапараметры для индексации в БЗ, для ускорения реализации и повышения эффективности поисковых запросов в глобальных сетях или выполнения заданий на анализ и обработку в пределах результативности рассматриваемой лингвистической БЗ.

 ISSN 1560-8956

Наиболее успешное применение метрики могут также найти для измерения или оценки состояний следующих компонентов решаемых задач:

1. Процессов анализа и обработки;
2. Полноты описаний лингвистических объектов (членов предложения, группы подлежащего или группы сказуемого, предложения, абзаца, страницы, т.п.);
3. Оценки отличий событий и фактов между собой – например, фактов хранящихся в БЗ и нового факта, полученного в виде конечного результата указанных выше процессов;
4. Оценки новых знаний для выявления их отличий от старых знаний, которые накоплены в БЗ на момент сравнения;
5. Повышения релевантности поисковых запросов при использовании оценок на основе введенных метрик описания лингвистических объектов.

Если ввести понятие состояния БЗ через набор ее информационных характеристик, то метрики состояния могут быть использованы для управления событиями в БЗ или для улучшения ее администрирования в целом.

Обобщая описанное выше, можно предложить, по аналогии с [4], метамодель метрик оценивания или измерения состояний объектов и процессов лингвистического анализа и обработки в следующем виде:

1. описание (название и краткая характеристика метрики)
2. спецификация (из чего она состоит и к каким объектам или процессам она относится)
3. обоснование (необходимость введения метрики)
4. условия и ограничения применения
5. алгоритм (или последовательность) ее вычисления или измерения
6. возможные размерности метрики (%, отношение A/B, т.п.)
7. параметры (целевое значение, среднее, опасное значение)
8. диапазон (min/ max)
9. возможные критерии оптимизации (E^2; | f |; т.д.)
10. контроль точности измерения
    - алгоритм поверки, т.е. контроль точности
    - приборы и шаблоны
    - max допустимая погрешность
    - периодичность или условия наступления поверки
11. аудитория (т.е., к кому из персонала относится метрика)

12. значение KPI (ключевой показатель эффективности)
13. уровень зрелости взаимодействия процессов (т.е. % используемых составных метрик относительно “суммы =составные +низовые”)
14. класс метрики (информационная / измеримая)
15. тип метрики (простая / составная / интегрированная).

Такая метамодель может служить шаблоном при формировании конкретных метрик.

Выводы

Показана возможность формирования различных метрик для измерения состояний различных процессов анализа и обработки лингвистических объектов текста - в пределах принятых ограничений относительно функциональной трактовки информации и знаний в разработанной БЗ для лингвистических приложений. Предложенные в статье метрики могут также использоваться для создания новых оценок полноты описаний таких объектов, измерения отличий событий и фактов между собой при анализе в лингвистической БЗ, выявления новых знаний в ней, как один из подходов повышения релевантности поисковых запросов в IT – среде, а также для повышения управляемости указанных процессов анализа и обработки в случае использования таких оценок на уровне метамоделей.

Литература

1. Разработка ПО: метрики программных проектов. Обзор ITC.UA, 2007, http://itc.ua/articles/ razrabotka_po_metriki_programmnyh _proektov_27774
2. Kagdi H., Collard M., Maletic J. Towards a Taxonomy of Approaches for Mining of Source Code Repositories // ACM SIGSOFT Software Engineering Notes, Proceedings of the 2005 international workshop on Mining software repositories MSR ’05. St. Louis, Missouri, 2005, P. 1–5.
3. А. А. Стенин, Ю. А. Тимошин, В. Г. Галаган, В.П. Ярченко. О функциональности базы знаний по анализу и обработке лингвистической информации. Адаптивні системи автоматичного управління - Дніпропетровськ: ДНВП Системні технології, 2012-Вип.21(41).-с.
4. Метрики управления IT- услугами / Питер Брукс; пер. с англ.- М.: Альпина Бизнес Букс, 2008.-283 с.
5. В.А. Молчанова, Т.А. Тумина. О формировании интегрированной базы знаний компании. – //Транспортное дело России, 2008.
6. Питер Джексон Введение в экспертные системы = Introduction to Expert Systems — 3-е изд. — М.: “Вильямс, 2001. — С. 624. —ISBN 0-201-87686-8
7. Шемсединов Т.Г. http://blog.meta-systems.com.ua/2011/01/blog-post_28.html

8. Баргесян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. 2-е издание. СПб: БХВ-Петербург, 2007. 375 с.

9. http://msquaredtechnologies.com/m2rsm/docs/rsm_metrics_narration.htm

Отримано 30.11.2012 р.