Технологическая инструкция для автоматизированного создания объектно-ориентированных баз знаний*

Научно-исследовательская компания «КуБ»

Версия 1.0 от 31.07.2009г.

Данная информация носит открытый характер и предназначена для ознакомления потенциальных лицензиатов с возможностями данной разработки и составом соответствующей ноу-хау.

НАЗНАЧЕНИЕ

Технологическая инструкция предназначена для автоматизации процесса создания объектно-ориентированных баз знаний. В результате выполнения инструкции возникает программный продукт — база знаний, включающая систему управления данными и средства отображения данных на экране компьютера. База знаний компилируется согласно инструкции в виде Веб-ориентированного приложения, которое устанавливается на выделенный сервер и обеспечивает работу нескольких пользователей посредством стандартных Интернет-браузеров.

ТЕРМИНОЛОГИЯ

Объекты — индивидуальные понятия предметной области, имеющие стандартные буквенно-цифровые обозначения.

Объектная ориентированность — свойство информационной системы, отражающее, что в ее составе хранятся или обрабатываются объекты.

Ассоциативная связь — нечеткое сопоставление одного объекта другому, выполняемое алгоритмическим методом, основанным на теоретических представлениях о взаимодействии объектов в составе целостной системы.

База знаний — база данных, в которой хранится информация об объектах, и система управления этой базой данных, основанная на ассоциативных взаимосвязях между объектами.

ОБЛАСТЬ ПРИМЕНЕНИЯ

Разработка позволяет создавать базы знаний в области наук о жизни: молекулярной биологии, нанобиотехнологий, молекулярной медицины, биотехнологии, биоинженерии и генной инженерии. Специфика области применения определяется наличием существенного количества объектов, которыми в молекулярной биологии являются наименования молекул, в том числе — генов, белков, лекарственных соединений, а также наличие обширных электронных библиотек текстовых описаний свойств данных объектов (примером такой библиотеки является Интернет-ресурс PubMed).

АКТУАЛЬНОСТЬ И НОВИЗНА РАЗРАБОТКИ

Информация о результатах научных экспериментов содержится в описательном виде в составе электронных библиотек статей и патентов. Количество описательных документов в области наук о жизни увеличивается вдвое каждые 4 года. Изложенная в научных статьях информация слабо структурирована, что приводит к низкой эффективности овладения новыми данными в ходе проведения научно-исследовательских работ. Кроме того, в связи с расшифровкой генома человека, возможности методов анализа биологических объектов существенно расширились. Обратной стороной возросшей эффективности инструментальных методов является необходимость сопоставлять с текущим уровнем знаний результаты системного характера, имеющие отношение к разным областям специализации.

Данная разработка позволяет в автоматическом режиме структурировать материал, содержащийся в научных описаниях, и представлять его в систематизированном виде в составе интерактивной информационной системы. Основным преимуществом разработки, по сравнению с аналогами, является полная автоматизация, тогда как в сходных информационных системах (например, GeneOntology) структурирование данных проводится экспертами. За счет автоматизации достигаются преимущества создаваемых баз знаний, обеспечивающие их конкурентоспособность:

- «несмещенная оценка», отражающая объективное состояние разработок в предметной области, а не совокупность субъективных экспертных мнений;
- повышение производительности и существенное снижение себестоимости. Как следствие — возможность постоянной актуализации входящих в базу знаний объектов и ассоциативных связей;
- возможность создавать специализированные базы знаний по разным направлениям научно-исследовательской деятельности, в том числе по направлениям,

*Технологическая инструкция охраняется в режиме коммерческой тайны. По вопросам лицензионного использования обращаться: www.ookub.ru.

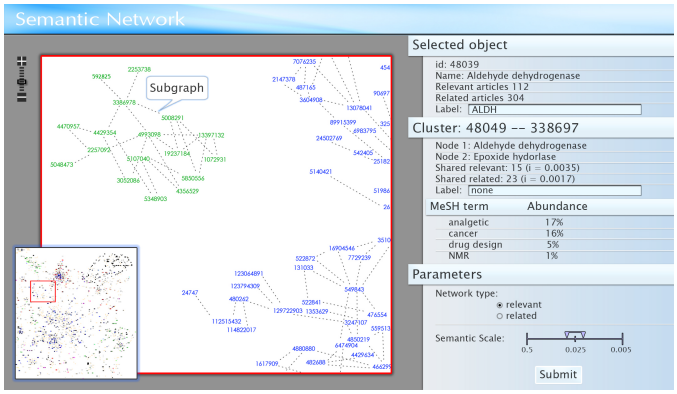

Рис. 1. Экран базы знаний, предназначенный для работы с семантической картой информационных объектов

имеющим интерес только для ограниченной группы пользователей.

Второй отличительной особенностью разработки является использование двух типов ассоциативных связей. Первый тип связи — тривиальный, он применяется во многих системах: PubGene, BioText, PubMatrix — связь устанавливается, когда наименования двух объектов содержатся в одном документе. Второй же тип связи устанавливается только в рамках данной ноу-хау — связь между объектами через расчет смыслового сходства документов. Преимуществом такого типа связи является возможность отображения в базе знаний скрытых взаимосвязей, которые могут стимулировать планирование медико-биологических экспериментов, направленных на получение патентно-чистых результатов.

РЕГЛАМЕНТИРОВАННЫЕ ПАРАМЕТРЫ

Время выполнения полного цикла работ по созданию базы знаний составляет не более 2-х недель, в том числе, в течение 1-й недели инструкция выполняется оператором с лаборантским уровнем квалификации, в течение 2-й недели производится просмотр базы знаний экспертом — кандидатом наук для удаления излишних объектов, инсталляцию осуществляет инженер-программист в течение 2-х часов. Общий объем рефератов, обрабатываемых в течение указанного периода, составляет от 100 до 150 тыс., в зависимости от области специализации базы знаний. Эксплуатация базы знаний может осуществляться на сервере разработчика в режиме подписки, на сервере заказчика или в портативном варианте – на нетбуке.

СОСТАВ РАЗРАБОТКИ

В состав технологической инструкции входят описания последовательности действий, обеспечивающие следующие этапы формирования базы знаний:

1. Сбор первичных данных от заказчика базы знаний, процедура интерактивной компьютерной обработки результатов анкетирования заказчика для конкретизации специальной области знаний.
2. Загрузка рефератов научных публикаций в соответствии со специализацией базы знаний, автоматическое распознавание в текстах загруженных документов наименований объектов.
3. Алгоритмическое определение ключевых дескрипторов, характеризующих свойства объектов.
4. Автоматический расчет оценки ассоциативной связи между объектами.
5. Распределение объектов по группам в соответствии со степенью ассоциативной связности.
6. Веб-ориентированная система управления базой данных, совмещенная с системой менеджмента Веб-контента.
7. Подготовка данных и их импорт в систему управления базой данных.
8. Система резервного копирования и восстановления данных.
9. Регламент актуализации данных в составе базы знаний (инструкция по ведению вспомогательного сайта технической поддержки).
10. Шаблон описания пользователя и методика его структуризации применительно к области специализации базы знаний.
11. Методика проведения тестовых испытаний очередного релиза базы знаний.
12. Инструкция для инсталляции базы знаний на сервер: под управлением ОС Linux, под управлением ОС Windows.

ЛИТЕРАТУРА

Jenssen TK, Laegreid A, Komorowski J, Hovig E. (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 28, 21-8.

Becker KG, Hosack DA, Dennis G Jr, Lempicki RA, Bright TJ, Cheadle C, Engel J. (2003) PubMatrix: a tool for multiplex literature mining. BMC Bioinformatics. 10, 4:61.

Homayouni et al. (2006) Semantic gene organizer (US Patent 20060047441)

Hearst MA, Divoli A, Guturu H, Ksikes A, Nakov P, Wooldridge MA, Ye J. (2007) BioText Search Engine: beyond abstract search. Bioinformatics. 23, 2196-7.

Plake C, Royer L, Winnenburg R, Hakenberg J, Schroeder M. (2009) GoGene: gene annotation in the fast lane. Nucleic Acids Res. 37, W300-4.

Пономаренко и соавт. (2009) Способ и компьютерная система для определения семантической связности между названиями химических соединений и биологических молекул (рег. №2009112450)

Таблица 1. Сравнение автоматических баз знаний (АБЗ) со сходными решениями

	АБЗ	GoGene	BioText	PubMatrix	PubGene
Типы объектов	Гены, белки, химические соединения	Гены, белки, биологические процессы, молекулярные механизмы	Нет	Любые термины	Гены, белки
Специализация	Да	Да	Нет	Нет	Нет
Автоматизация	Да	Нет	Да	Да	Да
Тип ассоциативности	Родственные, релевантные	Устанавливается экспертом	Родственные, релевантные	Релевантные	Релевантные
Лингвистический анализ	Нет	Нет	Да	Нет	Нет