Интеллектуальные информационные системы

Тема 6

Выделение терминов и связей между ними; тезаурусы

Павел Исаакович Браславский
pb@imach.uran.ru
весенний семестр 2006

План

- Введение
  - Что такое термин?
  - Что такое семантическая связь?
  - Что такое тезаурус? Примеры. Области приложения
  - Исходные данные для извлечения терминов и связей
- Обзор методов

Термин

Термины в области лексики и формула в области синтаксиса являются теми идеальными типами языкового выражения, к которым неизбежно стремится научный язык.

Шарль Балли

Определение

- **Термин** – слово или словосочетание, призванное точно обозначить понятие и его соотношение с др. понятиями в пределах специальной сферы. Т. служат специализирующими, ограничительными обозначениями характерных для этой сферы предметов, явлений, их свойств и отношений. Они существуют лишь в рамках определённой терминологии. В отличие от слов общего языка, Т. не связаны с контекстом. В пределах данной системы понятий Т. в идеале должен быть однозначным, систематичным, стилистически нейтральным. (БСЭ)

Свойства терминов

- Понятие ↔ Термин
- Устойчивость (повторяемость)
- Самостоятельность (не зависит от контекста)
- Эмоциональная нейтральность
- Взаимосвязанность → терминосистема
- Существование определения
- Лингвистические характеристики (обычно – именная группа)

Отношения между терминами

- Парадигматические vs Синтагматические отношения
- Примеры
  - Род-вид: мебель – стол – письменный стол
  - Часть-целое: автомобиль – двигатель – поршень
  - Комплимент: вилка – нож
  - Ассоциация: трактор – МТЗ

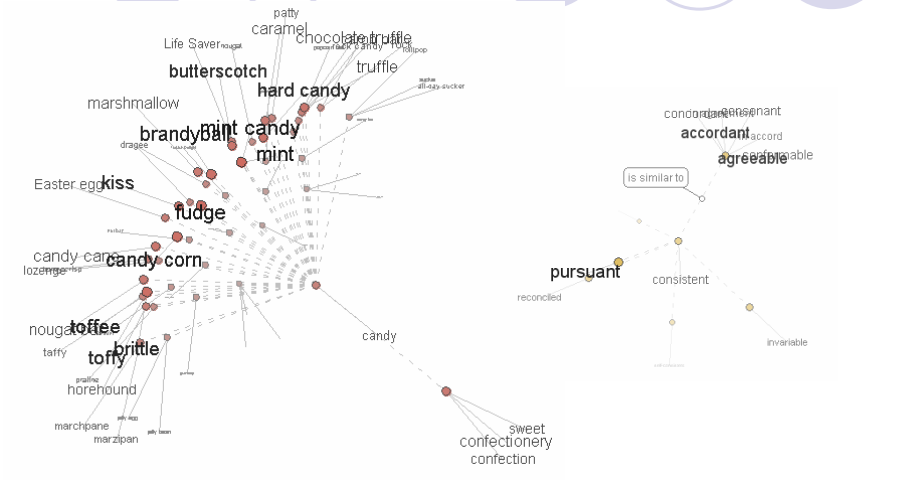
Тезаурус

- «Семантический словарь»: термины и связи между ними
- Онтологии (Semantic Web), таксономии
- Общеязыковые тезаурусы ↔ Тематические тезаурусы

Тезаурус: примеры

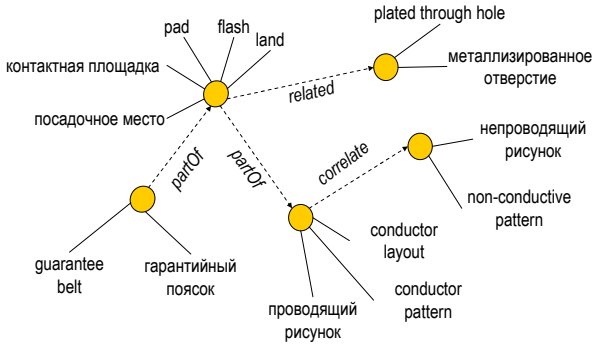
- Тезаурусы ЕЯ: тезаурус Роже, WordNet, RussNet, Euro WordNet, Russian WordNet
- РуТез
- ProThes: тезаурус «Автоматический оптический контроль печатных плат» (~200 понятий)

Visual Thesaurus (→WordNet)



www.visualthesaurus.com

AOI of PCB (ProThes)



Области приложения

- Компьютерная лексикография, терминоведение
- Машинный перевод
- Информационный поиск

Примеры использования

- Google Sets
- Google Synonym Search
- Google Definition Search
- Medical Subject Headings
- Библиотека УГТУ-УПИ
- ProThes

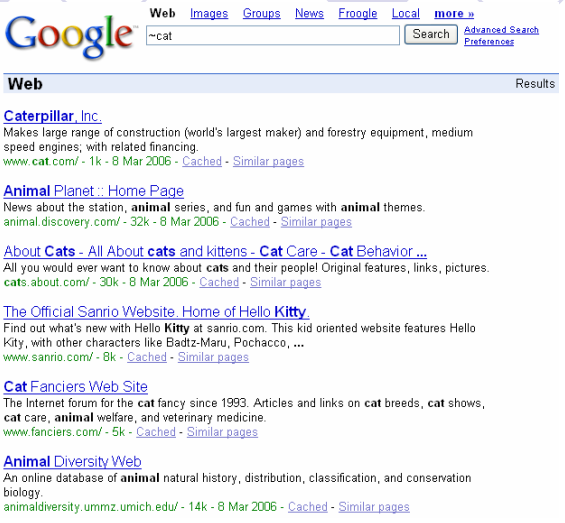
Google Sets

http://labs.google.com/sets

Mars → Mars, Venus, Mercury, Jupiter, Earth, Saturn, Missions, NASA, Moon, Neptune, Aerospace Engineering, Uranus, the Moon, Pluto

Mars, Snickers → Snickers, Mars, Twix, Bounty, Kit Kat, Double Decker, Cadbury's Fruit Nut, Maverik, Galaxy Caramel, Fuse, Yorkie Rasin Biscuit

Google Synonym Search



Google Definition Search

Web Images Groups News Froogle Local more »

Google define:risk management Search Advanced Search Preferences

Web

Related phrases: risk management plan risk management process risk management framework risk management zone risk management authority risk management risk analysis & risk management enterprise risk management financial risk management disaster risk management

Definitions of **risk management** on the Web:

- Decisions to accept exposure or to reduce vulnerabilities by either mitigating the risks or applying cost effective controls.
  www.utmb.edu/is/security/glossary.htm
- Decisions about whether an assessed risk is sufficiently high to present a public health concern and about the appropriate means for control of a risk significant. The process of evaluating and selecting alternative regulatory and non-regulatory responses to risk. The selection process necessarily req consideration of legal, economic, and behavioral factors.
  www.nsc.org/ehc/glossar2.htm
- The process of evaluating and selecting alternative regulatory and non-regulatory responses to risk. The selection process necessarily requires the co legal, economic, and behavioral factors.
  www.entrix.com/resources/glossary.aspx
- Risk management is the decision-making process involving considerations of political, social, economic and engineering factors with relevant risk ass relating to a potential hazard so as to develop, analyse and compare regulatory options and to select the optimal regulatory response for safety from t Essentially risk management is the combination of three steps: risk evaluation; emission and exposure control; risk monitoring.
  www.bio.hw.ac.uk/edintox/glossall.htm
- The identification and acceptance or offsetting of the risks threatening the profitability or existence of an organisation. With respect to foreign exchang among others consideration of market, sovereign, country, transfer, delivery, credit, and counterparty risk.
  www.fx-forex-trading.com/glossary.htm

Medical Subject Headings (MeSH)

National Library of Medicine - Medical Subject Headings

2006 MeSH

MeSH Supplementary Concept Da

Return to Entry Page

Name of Substance	phenazepam
Record Type	C
Registry Number	51753-57-2
Related Number	70030-06-7 (2-(14)C-labeled cpd)
Entry Term	7-bromo-5-(2-chlorophenyl) 1,3-dihydr benzodiazepin-2-one
Entry Term	7-bromo-5-(2-chlorphenyl)-1,2-dihydro benzodiazepin-2-one
Entry Term	fenazepam
Entry Term	phenazepam, 2-(14)C-labeled cpd
Heading Mapped to	*Benzodiazepines

National Library of Medicine - Medical Subject Headings

2006 MeSH

MeSH Descriptor Data

Return to Entry Page

MeSH Heading	Influenza in Birds
Tree Number	C02.782.620.375
Tree Number	C22.131.450
Tree Number	C22.131.728.450
Annotation	coordinate IM with specific virus subtype (IM) if pertinent; don't forget also BIRDS or POULTRY (NIM) & check tag ANIMALS
Scope Note	Infection of domestic and wild fowl and other BIRDS with INFLUENZA A VIRUS. Avian influenza usually does not

Поиск по рубрикатору library.ustu.ru

Программирование ЭВМ

см. раздел выше

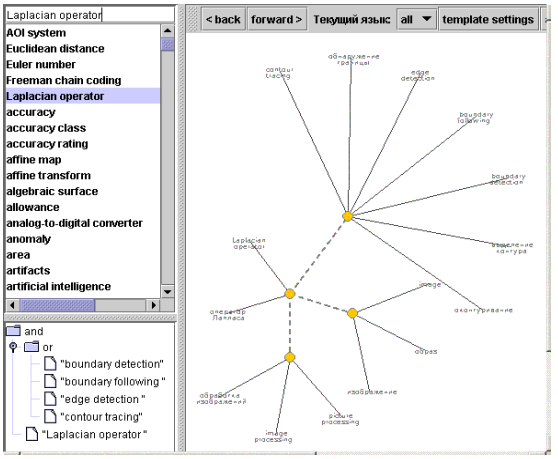
развернуть список операций

УДК: 004.42

- Надежность программ
- Типы данных
- Методы сортировки

Адресное программирование - Вычислительные машины (абоненские пункты) - Вычислительные машины (математическое обеспечен
Вычислительные машины (программирование) - Вычислительные машины (фонды алгоритмов и программ) - Загрузка файлов (вычисли
техника) - Интегрированные системы программирования - Концептуальное программирование - Логическое программирование -
Макрогенерация - Макросы - Мастер (интегрированная система программирования) - Математическое обеспечение (цифровые
вычислительные машины) - Микро-ЭВМ (программное обеспечение) - Программирование (вычислительная техника) - Программирова
(цифровые вычислительные машины) - Программные прерывания (вычислительная техника) - Программы ЭВМ - Программы ЭВМ
(векторизация) - Программы ЭВМ (интеллектуализация) - Программы ЭВМ (качество) - Программы ЭВМ (надежность) - Программы Э
(оптимизация) - Программы ЭВМ (преобразование) - Программы ЭВМ (сопровождение) - Программы ЭВМ (спецификация) - Програм
ЭВМ (эксплуатация) - Рекурсивное программирование - Систематическое программирование - Следящее программирование - Структу
программирование - Теоретическое программирование - Файлы (загрузка) - Фонды алгоритмов - Форматы файлов - Функциональное
программирование - Цифровые вычислительные машины (математическое обеспечение) - Цифровые вычислительные машины (на
интегральных микросхемах) - Цифровые вычислительные машины (программирование) - Цифровые вычислительные машины
(программирование) - Цифровые вычислительные машины (программы)

ProThes



http://imach.uran.ru/prothes/

Методы

- Статистические ↔ Символьные (шаблоны)
- Термины или Отношения (отдельно) ↔ (Термины + отношения)
- Точность ↔ Полнота
- Автоматические ↔ автоматизированные

Первичные источники

- Словари
- Корпус текстов (разметка? категоризация?)
- Web (индекс МП)
- Существующий список терминов / тезаурус

LEXTER

- POS – разметка
- Нахождение максимальной именной группы («негативное» определение NP)
- Нахождение кандидатов в термины внутри максимальной NP
- Корпус 200 000 → 10 000 кандидатов (20 000 употреблений)

Bourigault, 1992

Термины + правила вариаций

- Правила тех типов: согласование (25), вставка (17), перестановка (31)
  - *surgical closure → surgical exploration and closure: surgical exploration*
  - *medullary carcinoma → medullary thyroid carcinoma: thyroid carcinoma*
  - *control center → center for disease control: disease control*
- Формирование концептуальных классов
- 12 717 → 5 080, 1 000 → 2 329
- Точность 84.6%

Jacquemin, 1996

Смешанный подход

- Выделение словосочетаний
- Морфологические шаблоны
- Правила для предъявления эксперту

200 Мб → 300 000 кандидатов → 28 000 терминов

- Итеративная «сборка» терминов-словосочетаний
- Использование существующего тезауруса

Добров и др., 2003

Выделение гипонимов

- Шаблоны:
  - частота встречаемости
  - однозначность
  - легкость распознавания
- Например:
  - *NP_0 such as {NP_1, NP_2,…(or|and)} NP_n*
  - *{NP_1, NP_2,…} and other NP_0*
- Машинное обучение (?)
- Другие отношения (?)

Hearst, 1992

cereals:	rice* wheat*
countries:	Cuba Vietnam France*
hydrocarbon:	ethylene
substances:	bromine* hydrogen*
protozoa:	paramecium
liqueurs:	anisette* absinthe*
rocks:	granite*
substances:	phosphorus* nitrogen*
species:	steatornis oilbirds
bivalves:	scallop*
fungi:	smuts* rusts*
fabrics:	acrylics* nylon* silk*
antibiotics:	ampicillin erythromycin*
institutions:	temples king
seabirds:	penguins albatross*
flatworms:	tapeworms planaria
amphibians:	frogs*
waterfowl:	ducks
legumes:	lentils* beans* nuts
organisms:	horsetails ferns mosses
rivers:	Sevier Carson Humboldt
fruit:	olives* grapes*
hydrocarbons:	benzene gasoline
ideologies:	liberalism conservatism

«Сырой» тезаурус

- Извлечение значимых слов, близких слов, ассоциированных глаголов, выражений, однокоренных слов.
- Токенизация, POS-разметка, синтаксический разбор, вычисление меры сходства существительных (*Jaccard measure*) через глаголы, прилагательные и другие существительные
- Словосочетания (*NPs*)
- Однокоренные слова (сравнение подстрок + положение в тексте)

cancer :: [255 contexts, frequency rank: 29] MED *Relat.* lesion, tumor; tissue, disease; carcinoma. *Vbs.* advance, disseminate. *Exp.* cancer patient (cf. survival time, joint deformity), cancer chemotherapy (cf. survival time, intra-arterial infusion), cancer cell (cf. human cell, year period). *Fam.* cancer-specific, cancerous.

Grefenstette, 2001

Выделение терминов из Веба

- Нахождение терминов, связанных с данным (*seed term*)
- Этапы:
  - Формирование корпуса предложений, содержащих данный термин (с использованием МП)
  - Извлечение терминов-кандидатов
  - Фильтрация (частота, отношения и связи)
- Точность 85%, но низкая полнота

Sato and Sasaki, 2003

Семантическая близость

- Нет четкого определения в общем случае:
  - бегемот, гиппопотам
  - вилка, нож, ложка
  - танк, гитара
  - самолет, муха
  - инфекционный, дезинфицировать, антибиотик
- Контекстная взаимозаменяемость

Семантическая близость 1

Матрица документ-термин

	кошка	собака	трактор	ракета	спутник
док 1	2	3	0	0	0
док 2	0	2	0	1	1
док 3	0	0	1	0	0
док 4	0	0	0	2	4
док 5	0	0	1	1	1

Семантическая близость 2

Матрица термин-термин

	кошка	собака	трактор	ракета	спутник
кошка	2	3	0	0	0
собака	3	1	0	0	1
трактор	0	0	1	0	0
ракета	0	0	0	2	4
спутник	0	1	0	4	1

$B=A^TA$

Семантическая близость 3

Матрица термин-определитель

	кошка	собака	трактор	ракета	спутник
бешенный	0	2	0	0	0
любимый	2	1	0	0	0
космический	0	0	0	1	1
искусственный	0	0	0	0	3
мощный	0	0	5	2	0

Бинарные меры близости

$$Dice \quad \frac{2|X \cap Y|}{|X|+|Y|}$$

$$Jaccard \quad \frac{|X \cap Y|}{|X \cup Y|}$$

$$overlap \quad \frac{|X \cap Y|}{\min(|X|,|Y|)}$$

$$\cos \quad \frac{|X \cap Y|}{\sqrt{|X||Y|}}$$

Векторные варианты

$$Dice \quad \frac{\sum x_i y_i}{\sum (x_i + y_i)}$$

$$Jaccard \quad \frac{\sum \min(x_i, y_i)}{\sum \max(x_i, y_i)}$$

$$\cos \quad \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

Весовые функции

$$\frac{\log_2(x+1)}{\log_2(n+1)}$$

$$\log(x)+1$$