Автоматическое построение графа цитирований Задачи и методы

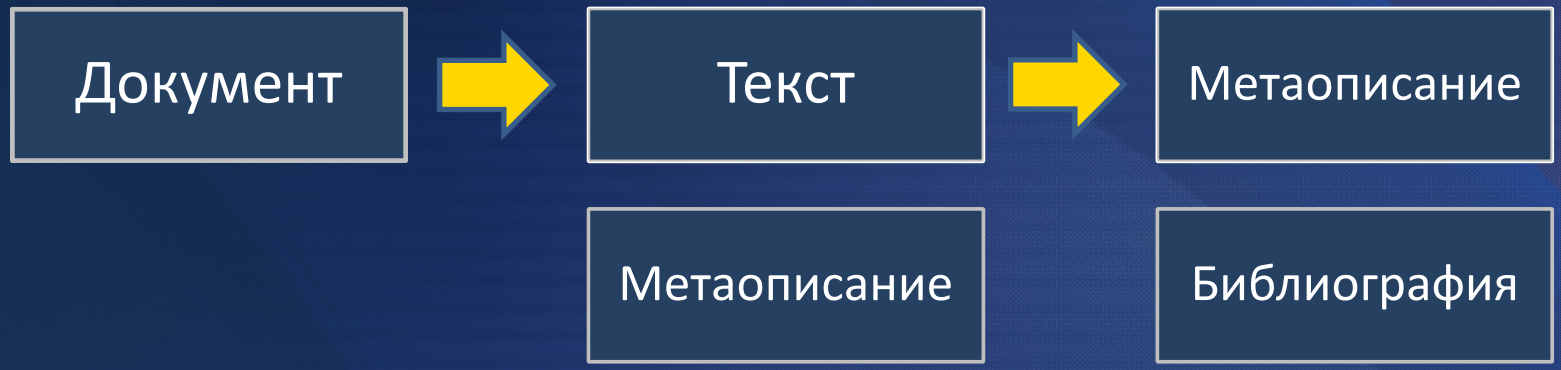
Полежаев Валентин

3 октября 2012

Документ. Граф цитирований. Представление данных.
Применение методов машинного обучения.

ВВЕДЕНИЕ

- Основные «единицы» документа



- Построить граф цитирований – связать библиографические записи одних документов с метаописаниями других.

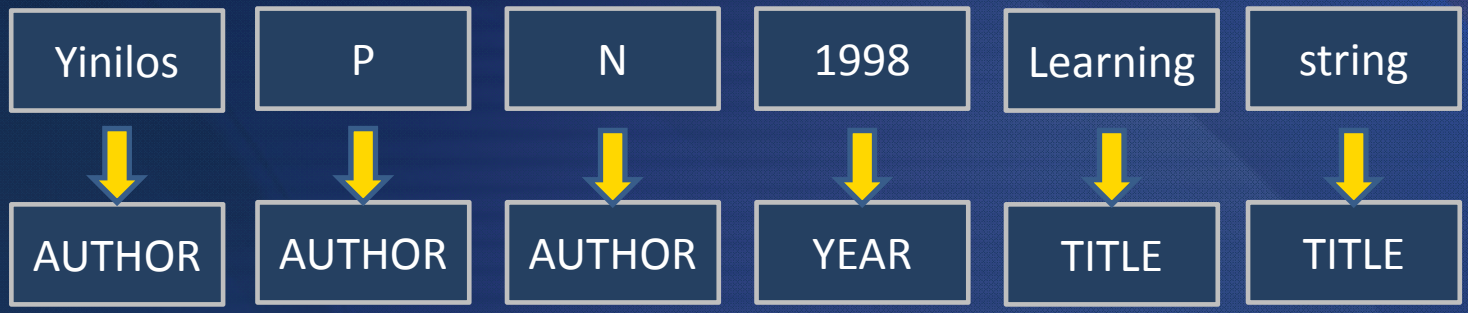
- Сложности
  - Большие размеры коллекций (миллионы документов)
  - Разнообразие оформлений текстов
  - Проблемы, связанные с извлечением текстов

Разметка. Сегментация. Связывание

ЗАДАЧИ АНАЛИЗА ТЕКСТОВ

Задача разметки

- Задача разметки заключается в построении поэлементного отображения последовательности входных символов на последовательность выходных символов



- Стандартные методы классификации
  - Объект классификации – элемент последовательности
  - Признаковое описание расширяется контекстными признаками
  - Сложности при учете классов соседних объектов
  - Ограниченность при учете зависимостей
- Лучший подход – использование графических моделей

Conditional Random Fields (CRF)

- Линейная модель CRF

$$P(X,Y) = \frac{1}{Z} F(X,Y) = \frac{1}{Z} \prod_{j=1}^{n} \prod_{i=1}^{m} \exp\{\mu_i \cdot f_i(y_{j-1}, y_j, X, j)\}$$

Например

$$f_i(y_{j-1}, y_j, X, j) = [y_{j-1} = \dot{y}][y_j = \ddot{y}]$$

$$f_i(y_{j-1}, y_j, X, j) = [x_j = \dot{x}][y_j = \dot{y}]$$

$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{1}{Z} F(X,Y) \cdot \frac{Z}{\sum_Y F(X,Y)} = \frac{F(X,Y)}{\sum_Y F(X,Y)}$$

- Основные черты
  - Максимально простые зависимости внутри Y
  - Сложные зависимости на X, полноценное признаковое описание
  - Не нужно считать Z, а $\sum_Y F(X,Y)$, $argmax_Y P(Y|X)$ считаются легко

Сегментация и связывание

- Сегментация похожа на разметку, но
  - Последовательности имеют характерную блочную структуру
  - Используются техники близкие к кластеризации
  - Часто основаны на «качественном» составе текста, текст рассматривается как «мешок слов»

- Связывание (идентификация, мэтчинг) отвечает на вопрос являются ли два объекта представлением одной сущности

AUTHOR: Eric Sven Ristad & Peter N. Yianilos **TITLE**: «Learning string edit distance» **…**	**AUTHOR**: Ristad, E. S., and Yianilos, P. N **TITLE**: Learning string edit distance **…**

$= ?$

Связывание. Стандартный подход

- Стандартный подход
  - Даны два объекта-структуры с текстовыми полями
  - Имеем N строковых полей и K функций сравнения двух строк
  - Строится соответствующий вектор сравнения размерности N·K
  - Решение об идентичности принимается бинарным классификатором (SVM) по полученному вектору сравнения

- Функции сравнения строк
  - Character-based. Расстояние Левенштейна, функция Джаро, …
  - Token-based. TF-IDF, функция схожести Жаккарда, …
  - Hybrid. Soft-TF-IDF, функция Монжа-Элкана, ...
  - Предикаты. «В строках одинаковое число слов», …

Существуют еще обучаемые функции, но на практике используются редко.

- Основная проблема – вычислительная сложность. В самом простом случае связывание на множестве объектов решается перебором всех пар.

Метаописание документа. Разметка. Классификация на строках.

ВЫДЕЛЕНИЕ МЕТАОПИСАНИЯ

Выделение метаописания

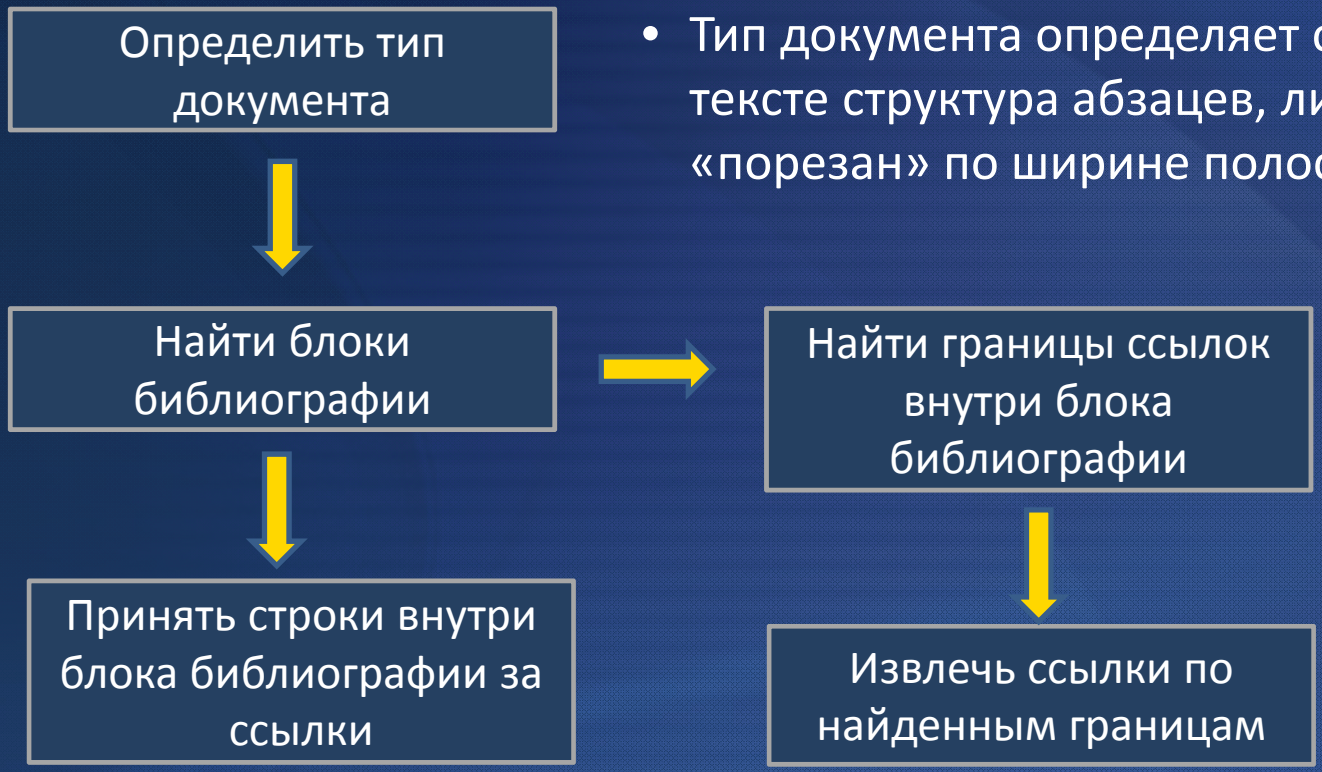
- Как задача разметки
  - Объекты – строки документа, классы соответствуют полям.
  - Для каждого поля свой бинарный классификатор
  - Классификация итерациями
  - Если строка отнесена к нескольким классам – эвристический алгоритм выделения частей в строке

- Признаковое описание
  - Характеристики строк (номер строки, число слов в строке, доля цифр и т.д.).
  - Словарь для каждого из полей.

Библиографические ссылки. Сегментация. Выделение блоков библиографии. Разметка. Определение ссылок внутри блоков.

ВЫДЕЛЕНИЕ БИБЛИОГРАФИЧЕСКИХ ССЫЛОК

Задача выделения библиографических ссылок



- Тип документа определяет сохранена ли в тексте структура абзацев, либо текст «порезан» по ширине полосы.

Определение типа текста

- Пример ссылки, «порезанной» по ширине колонки

> Mohri, M. (2000). Minimization algorithms for Sequential transducers. Theoretical Computer Science, 234, 177-201.

Вторая строка слабо отличима от обычного текста документа вне контекста соседних строк.

- Тип определяется решающим правилом по величине дисперсии длин строк в тексте

Выделение блока библиографии и ссылок внутри него

- Алгоритм выделения блока библиографии
  - Задача сегментации
  - Объекты – строки, классы – принадлежность блоку библиографии.
  - Решение основано на базовой (дерево C4.5) и контекстной (на результатах базовой) классификаций
  - Признаками являются характеристики строк (например, число заглавных букв)
  - Для учета контекста используется скользящее окно, либо суммирование значений соседних признаков

- Алгоритм выделения ссылок из блока библиографии
  - Задача разметки
  - Объекты – строки, классы – первая часть ссылки, либо нет.
  - Решается классификатором (дерево C4.5)

Библиографические ссылки. Разметка. CRF. FLUX-CiM.

СТАНДАРТИЗАЦИЯ

Применение Conditional Random Fields

- Задача разметки
  - По выделенным строкам библиографических записей получить стандартное представление в виде структур с текстовыми полями
  - Элементы – слова и разделители, классы соответствуют полям.
- Применение CRF
  - Модель линейная, либо Skip-chain.
  - Второго порядка, значение наблюдаемой переменной x_j зависит от значений скрытых переменных y_{j-2}, y_{j-1}, y_j.
  - Контекстные признаки, значение наблюдаемой переменной x_j зависит от значений соседних наблюдаемых переменных.
  - Использование словарей собственных имен при формировании признаков
- Особенности
  - Как правило, supervised, требует вручную размеченных данных.

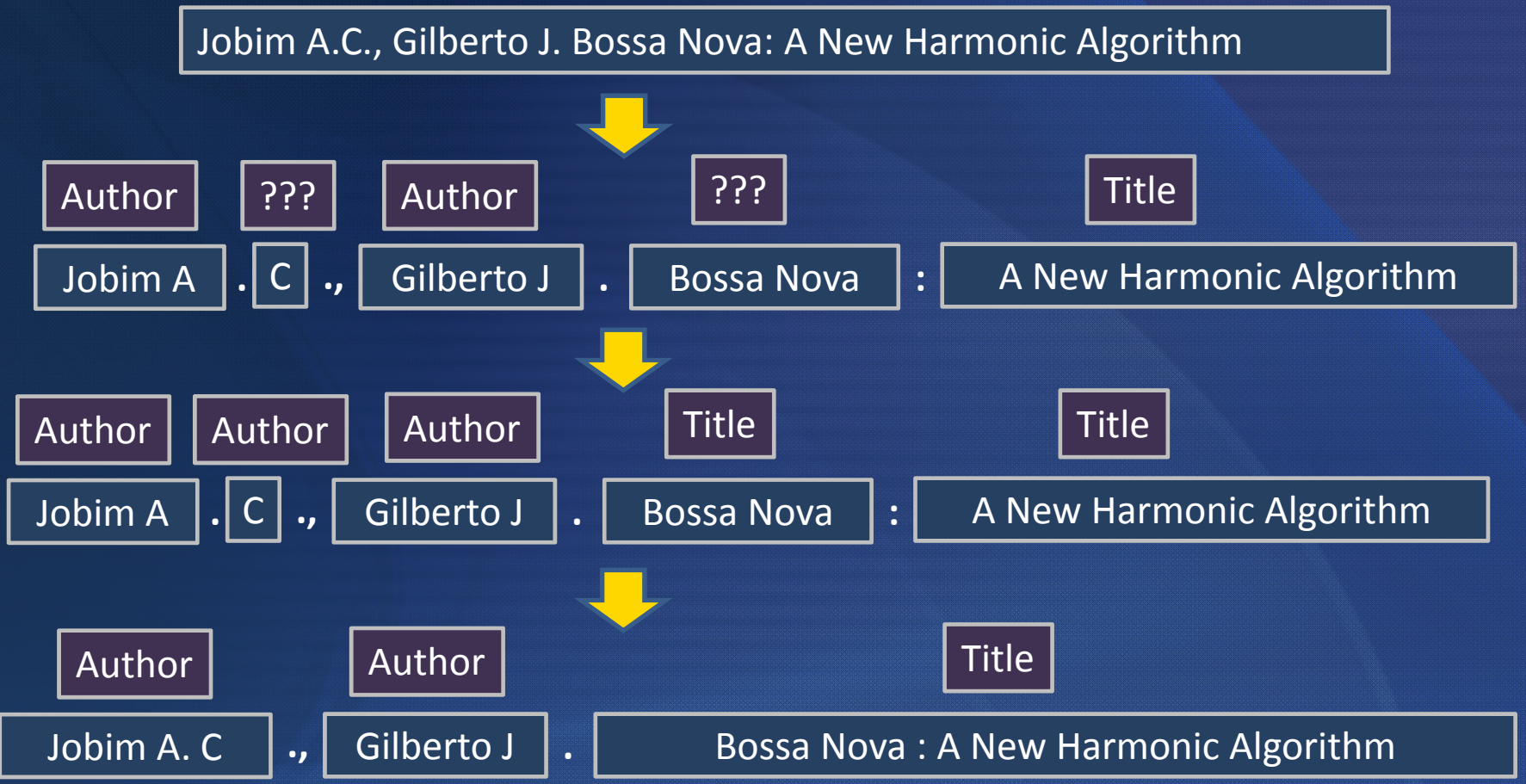
- Альтернатива CRF
  - Unsupervised, не требует размеченных вручную последовательностей
  - Построен на эвристиках

- Главный объект – «база знаний» (knowledge base):

AUTHOR:	**TITLE**:	**PUBLISHER**:	…
Monge	String	Springer	…
Pasula	Learning	Morgan Kauffman	
Ristad	Distance		
Cohen	Rules		
…	…	…	

Собирается автоматически из любых источников, предоставляющих структурированные данные (например DBLP)

FLUX-CiM. Пример работы



Критична полнота «базы знаний»

Адаптивный мэтчинг. Canopies. Основной алгоритм.

ПОСТРОЕНИЕ ГРАФА ЦИТИРОВАНИЙ

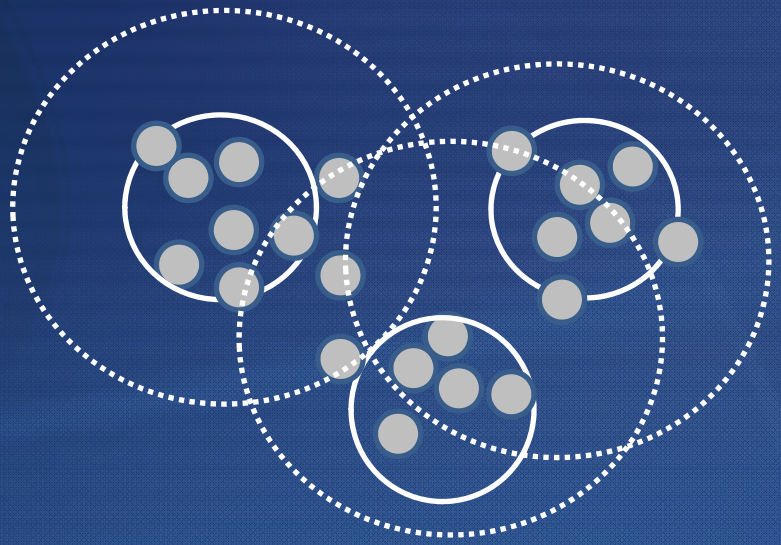
Задача связывания и Canopies

- Задача – связать метаописания и библиографические записи.

В обеих случаях это структуры с текстовыми полями. Работает стандартный способ связывания.

- Проблема в большом числе пар для связывания.

Для этого предлагается отбросить пары заведомо не идентичных объектов.



Canopies

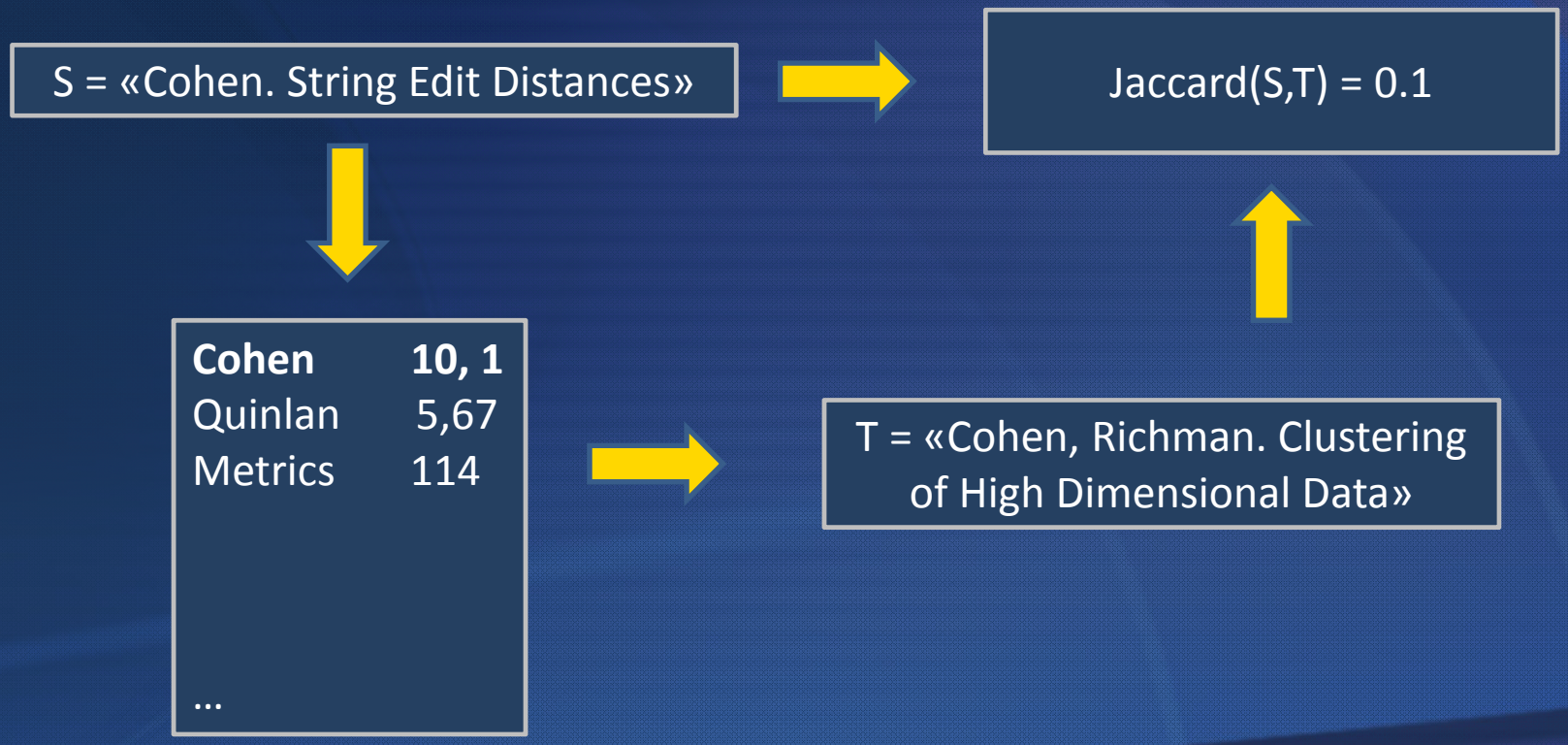
Canopies. Некоторые факты

- Некоторые факты:
  - Canopy – «слабый» кластер. Каждый объект может принадлежать нескольким canopy.
  - Непопадание объектов в один canopy интерпретируется как максимальная удаленность.
  - При довольно слабых ограничениях не изменяет результата кластеризации, если используется перед применением EM или алгоритма агломеративной кластеризации.
- Организация применительно к библиографическим записям
  - В качестве «дешевой» функции сравнения можно выбрать любую Token-based функцию.
  - Тогда легко организуется с помощью инвертированного индекса. В качестве ключей выступают слова (токены), в качестве значений – идентификаторы объектов (записей).
- Случай двух множеств
  - Алгоритм легко переносится на случай когда сравнивать объекты нужно не внутри некоторого множества (кластеризация), а между двумя заданными множествами (мэтчинг).

Canopies. Пример организации вычисления «дешевой» функции сравнения строк

Строки S,T, как «мешки слов»

$$\text{Jaccard(S,T)} = \frac{|S \cap T|}{|S \cup T|}$$



Адаптивный мэтчинг

- Процедура мэтчинга

  На входе: пара множеств для мэтчинга A,B, обученный бинарный классификатор (SVM).

  1. Получить пары-кандидаты методом Canopies
  2. Для каждой пары выполнить идентификацию стандартным способом (вектор сравнений, SVM), в результате будут получены компоненты связности
  3. Все объекты внутри одной компоненты считать идентичными

- Процедура обучения

  На входе: множества объектов A,B, пары, отмеченные как идентичные

  1. Получить пары-кандидаты методом Canopies
  2. Сформировать обучающую выборку из выбранных пар-кандидатов с учетом отмеченных пар
  3. Обучить классификатор

Окончательный алгоритм

- Обозначения

  O – метаописания в базе, S – не идентифицированные библиографические записи в базе, N – новые библиографические записи, I – новые метаописания, AM(A,B) – адаптивный мэтчинг на множествах A,B.

- Алгоритм обновления графа цитирований

  1. AM(O,I)

     Проверка наличия в базе добавляемых документов.

  2. AM(O∪S,N)

     Связываются новые записи либо с метаописаниями в базе, либо с находящимися в базе библиографическими записями.

  3. AM(I,S)

     Связываются новые метаописания с находящимися в базе библиографическими записями.