Выявление дубликатов в библиографических базах данных

Князева А.А., Турчановский И.Ю., Колобов О.С.

Институт вычислительных технологий СО РАН
Институт сильноточной электроники СО РАН

Частные задачи

- предварительная подготовка записей;
- способ составления пар из записей;
- способы сравнения отдельных полей записей;
- принятие решения о соответствии на уровне записи на основе результатов сопоставления отдельных полей.

aknjazeva@ict.nsc.ru

Требования к системе выявления дубликатов

- Отсутствие предположений о функции распределения признаков;
- Отказ от эмпирических правил для принятия решения о соответствии записей;
- Отсутствие требования независимости сравнительных признаков;
- Работа с записями в форматах семейства MARC;
- Возможность работы с неполными данными.

aknjazeva@ict.nsc.ru

Библиографическая запись

- элемент библиографической информации, фиксирующий в документальной форме сведения о документе, позволяющие его идентифицировать, раскрыть его состав и содержание в целях библиографического поиска (ГОСТ 7.76).

aknjazeva@ict.nsc.ru

Фрагмент библиографической записи

001 61/Н340-682478

200 1 $a Гетероморфизм ядер и цитогенетические нарушения в бинуклеарных Т-лимфоцитах человека по влиянием вируса Эпштейна-Барр

700 1$a Шилов $b Б. В.$g Борис Владимирович$c цитолог $f 19710323

$3 AShilov_BoriB20031006634800700

71202 $a Сибирский медицинский университет $c Томск

6061#$a ГЕРПЕСВИРУС 4 ЧЕЛОВЕКА $x патогенность $3 D004854Q000472 $2 mesh $8 rus

6061#$a ЧЕЛОВЕК $x HUMAN

aknjazeva@ict.nsc.ru

Модель выявления дубликатов

$$M = \langle \alpha(a), \beta(b) \rangle; a = b; \alpha(a) \in A; \beta(b) \in A$$

$$U = \langle \alpha(a), \beta(b) \rangle; a \neq b; \alpha(a) \in A; \beta(b) \in A$$

$$x_i(\alpha, \beta) = X_i, i = 1, \ldots, K \qquad \gamma = (X_1, \ldots, X_K)^T$$

$$\Gamma^M = \{\gamma[\alpha(a), \beta(b)] | \langle \alpha(a), \beta(b) \rangle \in M\}$$

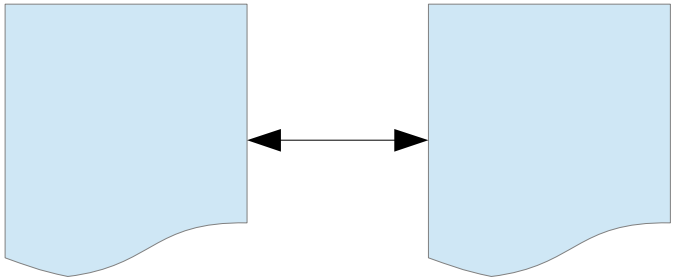
$$D(\gamma[\alpha(a), \beta(b)]) = \begin{Bmatrix} 1, \alpha(a), \beta(b) \in M, \\ 0, \alpha(a), \beta(b) \in U \end{Bmatrix}$$

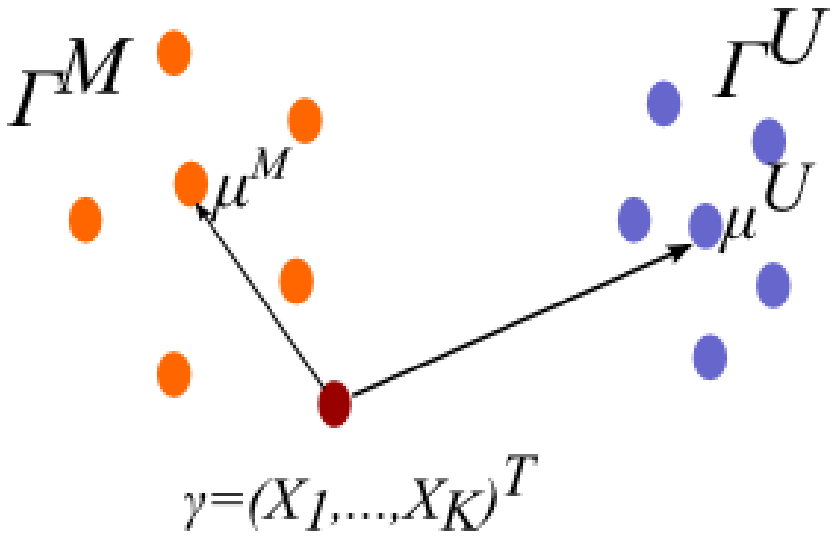
$$min \sum_{i=1}^{N} I\{D(\gamma_i[\alpha(a), \beta(b)]) \neq s(a, b)\}$$

$$Dist^2(\gamma, \mu_M) = (\gamma - \mu_M) W^{-1} (\gamma - \mu_M)^T$$

aknjazeva@ict.nsc.ru

Классификация пар записей





Вычисление сравнительных характеристик:

$$\gamma=(X_1, \ldots X_k)^T$$

aknjazeva@ict.nsc.ru

Список признаков

Признак	Описание	Сравнение	Поле
out	Соответствие по ISBN	Точное совпадение	010$a
title	Заглавие	Нечеткое сравнение	200$a
authors	Автор	Нечеткое сравнение	700$a, 700$b, 701$a, 701$b
place	Место издания	Нечеткое сравнение	210$a

aknjazeva@ict.nsc.ru

Список признаков

Признак	Описание	Сравнение	Поле
year	Год издания	Точное совпадение	210$d
publisher	Издатель	Нечеткое сравнение	210$c
edition	Сведения об издании	Точное совпадение номера	205$a
pages	Количество страниц	Точное совпадение	215$a
links	Коды связанных документов	Точное совпадение	423, 461

aknjazeva@ict.nsc.ru

Заключение

Особенности предлагаемого подхода:

- Возможность обучения;
- Работа с неполными данными;
- Возможность учета взаимосвязанных признаков.

aknjazeva@ict.nsc.ru

Спасибо за внимание!

aknjazeva@ict.nsc.ru

Выявление дубликатов в библиографических базах данных

Князева Анна

aknjazeva@ict.nsc.ru