Применение теории нечетких множеств к задаче генеалогической классификации текстов списков средневекового произведения (на примере двух древнерусских письменных памятников)

Семинар по социофизике
имени Д.С. Чернавского,
Институт проблем управления РАН
Москва, 22.10.2019

Актуальность темы исследования

- Совершенствование методов исследования
- Привлечение математических методов
- Адаптация теории нечетких множеств к историко-текстологическим исследованиям
- Универсальный и адекватный инструмент

Предмет и цель исследования

·Разработка новой методологии и метода количественной формализованной классификации средневековых текстов с большой рукописной традицией

·Проведение классификации двух средневековых произведений

Объект исследования

- Закон Судный людем (57 списков)
- Предсловие покаянию (21 список)

План доклада

- Элементы текстологии
- Применение мат. Методов в текстологии
- Модель нечеткой классификации
- Элементы теории нечетких множеств

План доклада

- Алгоритм нечеткой классификации
- Результаты классификации - стеммы
- Проблемы и задачи

Задачи

- Реконструкция архетипа
- Тексты исторических источников имеют свою историю
- Деление на текстологически близкие группы

Традиционный анализ

- Реконструкция архетипа
- Тексты исторических источников имеют свою историю
- Деление на текстологически близкие группы

Традиционный анализ

- Виды - незначительные, количественные различия
- Изводы – языковая специфика
- Редакции – целенаправленная переработка текста

Традиционный анализ. Виды ошибок по Д.С. Лихачеву

- Прочтения
- Запоминания
- Внутреннего диктанта
- Письма
- Переосмысления
- Стилистические изменения
- Идейные изменения
- Глоссы и интерполяции
- Работа писца по нескольким оригиналам - правка

Традиционный анализ. Примеры

- бо господа та -> богатства
- всоудъ да възьмоуть -> всегда возмоуть
- сеи соудъ-> съсоудъ
- иже коупетроу-> иже кусестру / иже петру южику
- О полоне-> о полhнh

Традиционный анализ. Восстановление первоначального текста

- Сравнительный анализ текстов списков
- Общие ошибки – результат переписи текста с копии, уже содержавшей ошибки
- Группировка по наличию общих ошибок
- Построение генеалогического древа

Критика метода Лахмана

- Каждый переписчик пользуется одним источником (протографом)
- Одинаковые ошибки писцы делали независимо друг от друга
- “Раздвоенность” стеммы (Ж. Бедье)

Формализация как вспомогательное средство работы текстолога

- Верификация работы текстолога
- Возможность работы с большим количеством текстов списков
- Возможность косвенного датирования неизвестных списков по известным

Формализованная классификация текстов списков

- Метод групп (Д. Фроже, 60-е гг. XX в.; Л.И.Бородкин и соавторы, 70-е гг. XX в.)
- (Colwell, Tune, 1963), А. Деес, Э. Ваттель

- Матрицы расстояний (Д. Фроже)
- Матрицы близости (Л.И.Бородкин)

- Агрегированные модели (Л.И. Бородкин)

Направления

- Кластерные методы – группировка, внутри которых объекты тесно связаны друг с другом

- Кладистические – принцип парсиномии:

в процессе эволюции переход из оного состояния в другое должен произойти за минимальное число шагов

Матрица близости

- Чем “ближе” генеалогически пара списков друг другу, тем меньше различий содержат их тексты

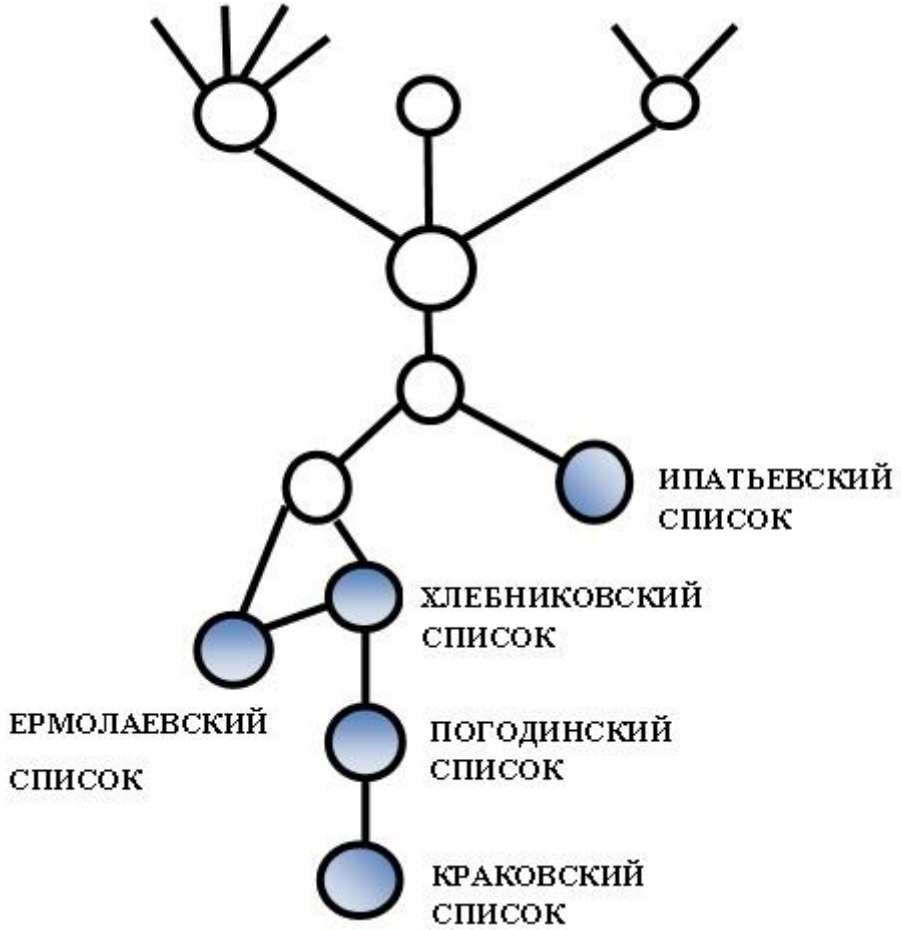
- Сличение списков: Число совпадающих слов

- Максимизация суммарного веса ребер графа связи между списками
- **Проблема**: Матрица симметрична!

Метод групп

- У каждого списка только один протограф
- В каждом списке - все ошибки протографа
- Нет одинаковых ошибок в списках, у которых независимые протографы
- **Проблема**: Жесткие требования к модели

Генеалогическая классификация

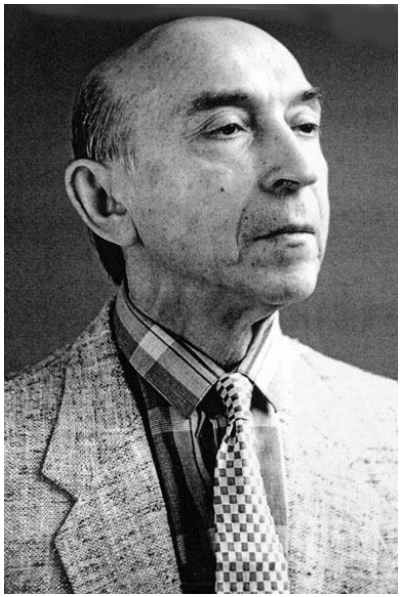


ФРАГМЕНТ СТЕММЫ ИПАТЬЕВСКОЙ ЛЕТОПИСИ (М.Д. ПРИСЕЛКОВ, 1940)

Модель нечеткой классификации

- Средневековый текст разнороден по составу
- Трудно указать основной список
- Можно установить предшествование одного другому лишь с определенным уровнем достоверности

Постановка задачи: Идея



Lotfi Askar Zadeh

род. 04.02.1921 г., Новханы, Азербайджанская ССР

- Fuzzy Sets. Information and Control. N 8. 1965; …
- Понятие лингвистической переменной и его применение к принятию приближенных решений. М., 1976.

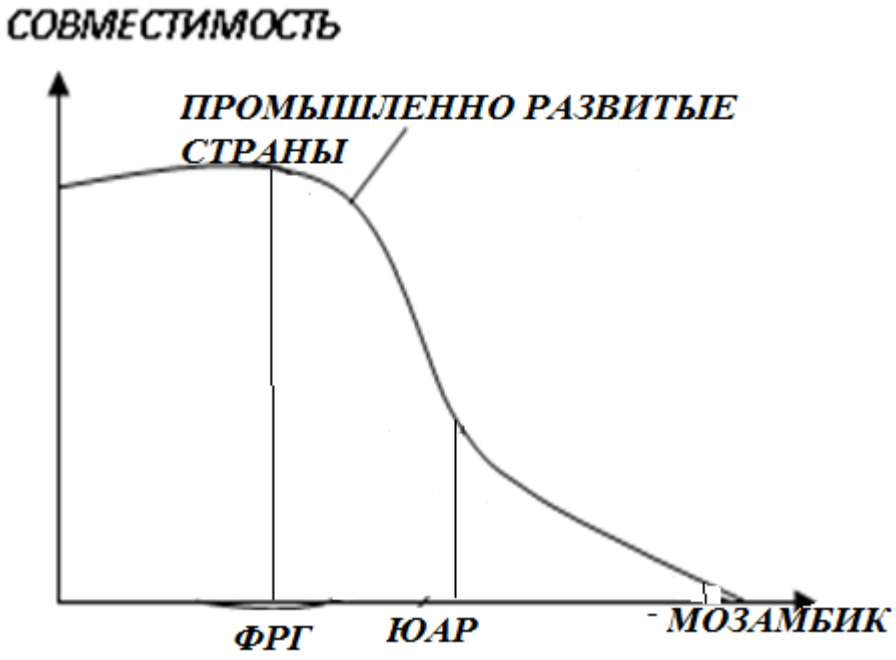
Теория нечетких множеств

- Принадлежность элементов множеству НЕ является однозначной
- Нет жестких границ между множествами
- Учет специфики задач классификации и типологии в гуманитарных науках

Нечеткие множества

- Нечеткое множество $\tilde{A}$ это множество пар

$$\tilde{A} = \{< \mu_A(x) / x >\}, \text{ где } x \in X,$$



Функция принадлежности множества $\tilde{A}$:

$$\mu_A(x) : X \to [0,1]$$

Нечеткое отношение

Определение *Нечетким отношением* на множестве X называется и через $\tilde{\eta}(X, V)$ обозначается пара множеств, в котором V является нечетким подмножеством на произведении $X \times X$: $V = \{\mu_V < x, y > / x, y \in X\}$, где функция принадлежности $\mu_V <\cdot,\cdot>$ ставит в соответствие любой паре элементов базового множества некоторое число из отрезка от 0 до 1: $\mu_V <\cdot,\cdot> \in [0,1]$

Нечеткие отношения: свойства

- Транзитивность
- Рефлексивность и антирефлексивность
- Симметричность и антисимметричность
- Связанность

Нечеткие отношения: свойства

- Эквивалентности:

  транзитивно, рефлексивно, симметрично;

- Порядка (совершенного строгого):

  транзитивно, антирефлексивно, антисимметрично и связанно

Нечеткие отношения: свойства

Определение 1. Отношение $\tilde{\eta} = (X, V)$ называется нечетко рефлексивным, если $\mu_V < x, x > \geq \delta$ для любого $x \in X$.

Определение 2. Отношение $\tilde{\eta} = (X, V)$ называется нечетко антирефлексивным, если $\overline{\mu_V < x, x >} \geq \delta$ для любого $x \in X$

Определение 3. Отношение $\tilde{\eta} = (X, V)$ называется нечетко симметричным, если $(\mu_V < x, y > \rightarrow \mu_V < y, x >) \geq \delta$ для любых $x, y \in X$, $x \neq y$.

А.Н. Мелихов, Л.С. Бернштейн, С.Я. Коровин.
Ситуационные советующие системы с нечеткой логикой. М., 1990. 272 с.

Нечеткие отношения: свойства

Определение 4. Отношение $\tilde{\eta} = (X, V)$ называется нечетко антисимметричным, если $\overline{\mu_V < x, y > \& \mu_V < y, x >} \geq \delta$ для любых $x, y \in X$, $x \neq y$.

Определение 5. Отношение $\tilde{\eta} = (X, V)$ называется нечетко транзитивным, если $\left(\bigcup_{y \in X}(\mu_V < x, y > \& \mu_V < y, z >) \rightarrow \mu_V < x, z >\right) \geq \delta$ для любых $x, y, z \in X,\ x \neq y, z \neq y$.

Определение 6. Отношение $\tilde{\eta} = (X, V)$ называется нечетко связанным, если $(\mu_V < x, y > \cup\, \mu_V < y, x >) \geq \delta$ для любых $x, y \in X,\ x \neq y$.

А.Н. Мелихов, Л.С. Бернштейн, С.Я. Коровин.
Ситуационные советующие системы с нечеткой логикой. М., 1990. 272 с.

Модель нечеткой классификации: Предположения

- Анализируем только тексты
- Есть текстолог, выделяющий нормальные чтения и уклонения от них (ошибки) в зависимости от времени (и места) создания списка – ранжирование в хронологическом порядке

- Чем больше ошибок одного списка есть в другом, тем достовернее, что первый список предшествует второму

- Чем меньше общих ошибок, тем более “независимы” списки друг от друга

Пример построения отношения предшествования

N места	Список x	Список y	Общие ошибки списков x и y S(x,y)	Ошибки списка x S(x)	Ошибки списка y S(y)
1	норма	норма	0	0	0
2	норма	ошибка	0	0	1
3	ошибка	ошибка	1	1	2
4	ошибка	ошибка’	1	2	3
5	ошибка	норма	1	3	3
6	норма	ошибка	1	3	4

Нормировка $V(x,y)=S(x,y)/S(x)=1/3$

$V(y,x)=S(x,y)/S(y)=1/4$

Учет различного влияния разночтений на классификацию

1. **Типологизация разночтений**

2. **Качественная шкала степеней значимости типов разночтений**

3. **Матрица попарных сравнений элементов типологического ряда**

4. **Определение весовых коэффициентов**

Метод косвенной экспертизы

Saaty T.L. Measuring the fuzziness of sets // Journal of Cybernetics. V. 4. 1974. Pp. 53-61.

Типологический ряд разночтений

1. **лексические замены**
2. **пропуски/ вставки глав**
3. **пропуск/ вставка предложений или их части**
4. **пропуск/ вставка словосочетаний**
5. **пропуск/ вставка слов**
6. **пропуск/ вставка частицы «не»**
7. **смысловые ошибки**
8. **пропуск/ изменение названия главы**
9. **перестановка слов/ нескольких слов**
10. **исправления**

Типологический ряд разночтений

11. **Орфография**
12. **описки, включая повторы**
13. **Русизмы**
14. **Архаика**

Качественная шкала степеней значимости типов разночтений (фрагмент)

Интенсивность значимости	Качественные оценки	Объяснения
0	Тип *i* несравним с типом *j*	Нет смысла сравнивать типы
1	Тип *i* одинаков по значимости с типом *j*	Типы равны по значимости
3	Тип *i* слабо значимее по сравнению с типом *j*	Существуют показания о предпочтении одного типа другому, но показания неубедительные
5	Тип *i* существенно или сильно	Существует хорошее доказательство и логические критерии, которые

Матрица попарных сравнений элементов типологического ряда

nn· ти-пов	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV¤
I	1¤	1¤	3¤	4¤	5¤	3¤	2¤	4¤	8¤	8¤	7¤	9¤	8¤	5¤
II	1¤	1¤	3¤	4¤	5¤	5¤	3¤	6¤	8¤	8¤	8¤	9¤	8¤	5¤
III	1/3¤	1/3¤	1¤	7¤	8¤	8¤	2¤	2¤	7¤	8¤	8¤	9¤	8¤	5¤
IV	¼¤	1/4¤	1/7¤	1¤	3¤	1/7¤	6¤	7¤	3¤	8¤	8¤	8¤	1/5¤	1/5¤
V	1/5¤	1/5¤	1/8¤	1/3¤	1¤	1/8¤	1/5¤	1/7¤	1¤	5¤	3¤	7¤	1/7¤	1/7¤
VI¤	1/3¤	1/5¤	1/8¤	7¤	8¤	1¤	1/7¤	0¤	1¤	3¤	0¤	3¤	1/8¤	1/8¤
VII¤	½¤	1/3¤	½¤	1/6¤	5¤	7¤	1¤	1/8¤	1/3¤	7¤	8¤	9¤	8¤	1/8¤

Вычисление весовых коэффициентов

$w(\cdot)$¤	**I**¤	**II**¤	**III**¤	**IV**¤	**V**¤	**VI**¤	**VII**¤	¤
Ненормализованные¤	0,1552¤	0,1705¤	0,1433¤	0,0819¤	0,0147¤	0,0443¤	0,0679¤	¤
Нормализованные¤	0,91¤	1¤	0,84¤	0,48¤	0,086¤	0,26¤	0,398¤	¤
$w(\cdot)$¤	**VIII**¤	**IX**¤	**X**¤	**XI**¤	**XII**¤	**XIII**¤	**XIV**¤	¤
Ненормализованные¤	0,116¤	0,0228¤	0,0072¤	0,0068¤	0,0065¤	0,0689¤	0,0941¤	¤
Нормализованные¤	0,68¤	0,134¤	0,042¤	0,04¤	0,038¤	0,404¤	0,552¤	¤

Матрица нечеткого отношения предпочтения (фрагмент) для «Закона Судного людем»

1·(Нв):·1,00·0,65·0,70·0,65·0,56·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·
0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,65·0,70·0,70·0,70·
0,70·0,70·0,70·0,70·0,70·0,70·0,70·0,70·0,70·0,70·0,70·0,77·0,77·0,77·0,77·0,57·0,53·0,72·0,77·
0,65·¶

¶

2·(Ч):·0,35·1,00·0,35·0,92·0,35·0,92·0,89·0,96·0,90·0,97·0,92·0,89·0,95·0,89·0,92·0,90·0,95·0,95·
0,96·0,89·0,89·0,89·0,89·0,89·0,89·0,89·0,89·0,89·0,89·0,87·0,87·0,89·0,93·0,89·0,35·0,35·0,35·
0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,37·0,37·0,37·0,37·0,35·0,35·0,35·0,35·
0,82·¶

¶

3·(ГII):·0,38·0,45·1,00·0,45·0,56·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·
0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,45·0,77·0,77·0,77·
0,77·0,77·0,77·0,77·0,77·0,77·0,77·0,77·0,77·0,77·0,77·0,44·0,44·0,44·0,44·0,38·0,38·0,38·0,69·
0,45·¶

¶

4·(З):·0,35·0,78·0,35·1,00·0,35·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·
0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,78·0,35·0,35·0,35·
0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,35·0,37·0,37·0,37·0,37·0,35·0,35·0,35·0,35·
0,78·¶

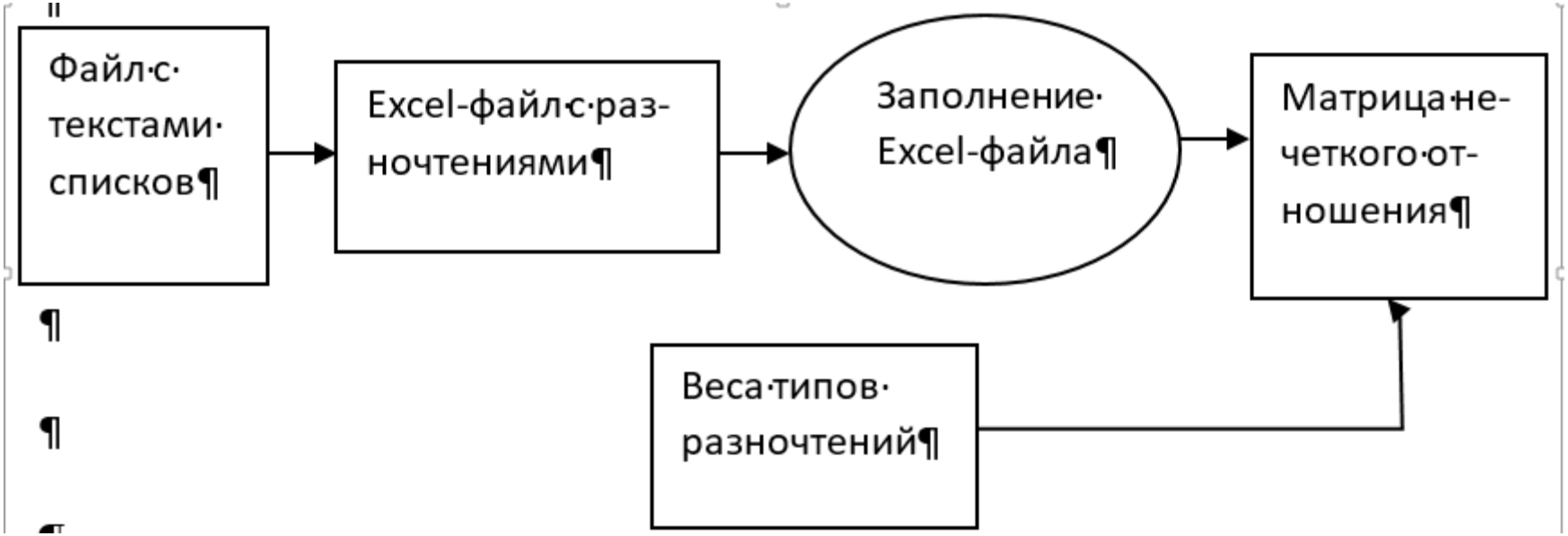
¶

5·(Пчт):·0,24·0,24·0,24·0,24·1,00·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·
0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·0,24·

Алгоритм нечеткой классификации текстов списков. Подготовительный этап

- Ввод текстов
- Выделение узлов разночтений
- Хронологическое ранжирование отдельных чтений

Алгоритм нечеткой классификации текстов списков. Подготовительный этап



Алгоритм нечеткой классификации текстов списков. Подготовительный этап

	A	B	C	D	E	F	G	H	I	J
1					Нв	Ч	ГП	З	Пчт	РМ
2	1	1		законъ*сѹдныи*людемъ						
3	2	2		правило*царя*константина						
4	2	1975		[]						
5	3	3		законъ-константина						
6	3	1976		[[]]						
7	3	5064		константина-законъ						
8	4	4		соѹдныи						
9	4	1977		сѹдныи						
10	4	4599		соѹдны						
11	4	6019		сѹд?ныи						
12	5	5		[]						
13	5	3782		всем						
14	6	6		правил?						
15	6	1978		[]						
16	7	7		ц(а)ря						
17	7	1979		[]						

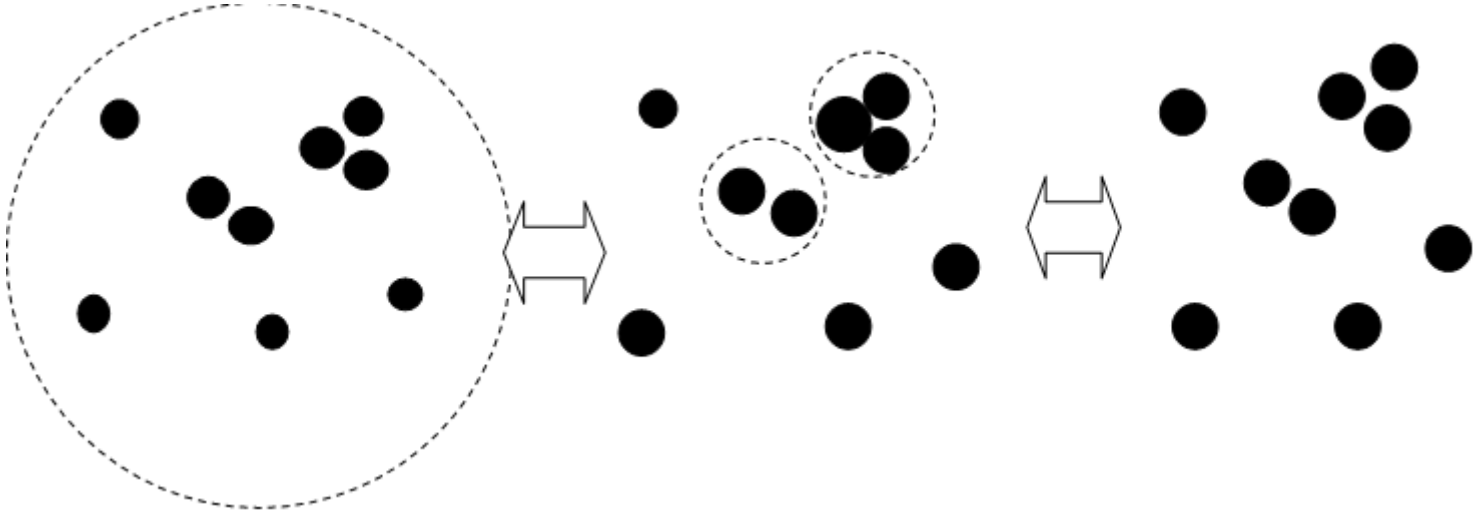
Алгоритм нечеткой классификации текстов списков. Подготовительный этап

	A	B	C	D	E	F	G	H	I
1					Нв	Ч	ГП	З	Пчт
2	1	1	0	законъ*сѹдныи*людемъ	0	0	0	0	0
3	2	2	8	правило*царя*константина	1	1	1	1	1
4	2	1975	8	[]	0	0	0	0	0
5	3	3	9	законъ-константина	1	1	1	1	1
6	3	1976	4	[[]]	0	0	0	0	0
7	3	5064	9	константина-законъ	2	2	2	2	2
8	4	4	11	соѹдныи	1	0	0	0	0
9	4	1977	11	сѹдныи	0	1	1	1	1
10	4	4599	11	соѹдны	1	0	0	0	0
11	4	6019	11	сѹд?ныи	0	1	1	1	1
12	5	5	5	[]	0	0	0	0	0
13	5	3782	5	всем	1	1	1	1	1
14	6	6	5	правил?	1	1	1	1	1
15	6	1978	5	[]	0	0	0	0	0
16	7	7	5	ц(а)ря	1	1	1	1	1
17	7	1979	5	[]	0	0	0	0	0
18	7	5065	5	ц(е)с(а)ря	1	1	1	1	1

Выделение текстологически близких списков

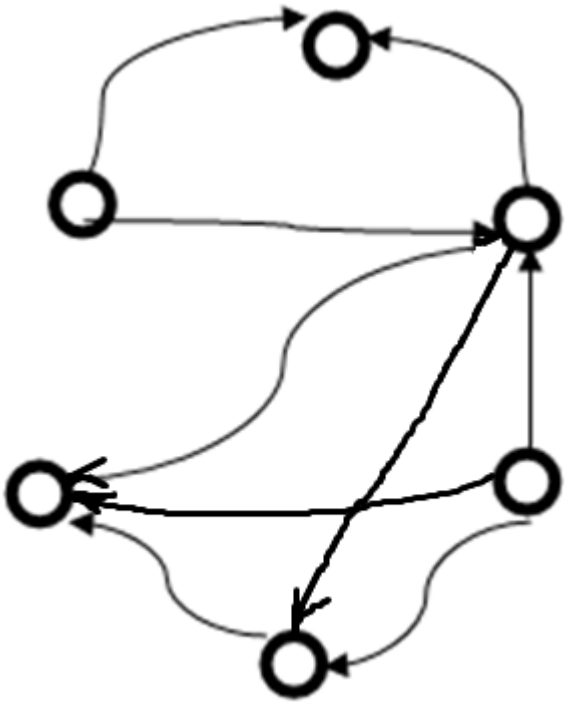
- Задание уровня достоверности.

- Разбиение списков на непересекающиеся эквивалентные классы



Хронологическое ранжирование текстов списков

- Выделение в каждом классе эталонного списка (текста)



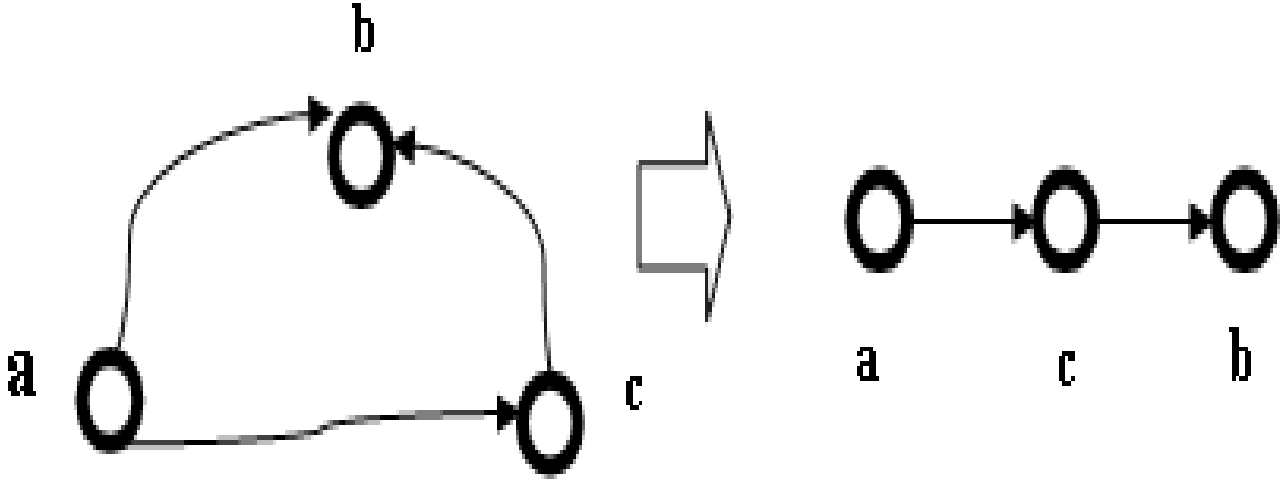
- Построение отношения порядка на множестве эталонов

- Выделение связных подмножеств на множестве эталонов

Хронологическое ранжирование текстов списков

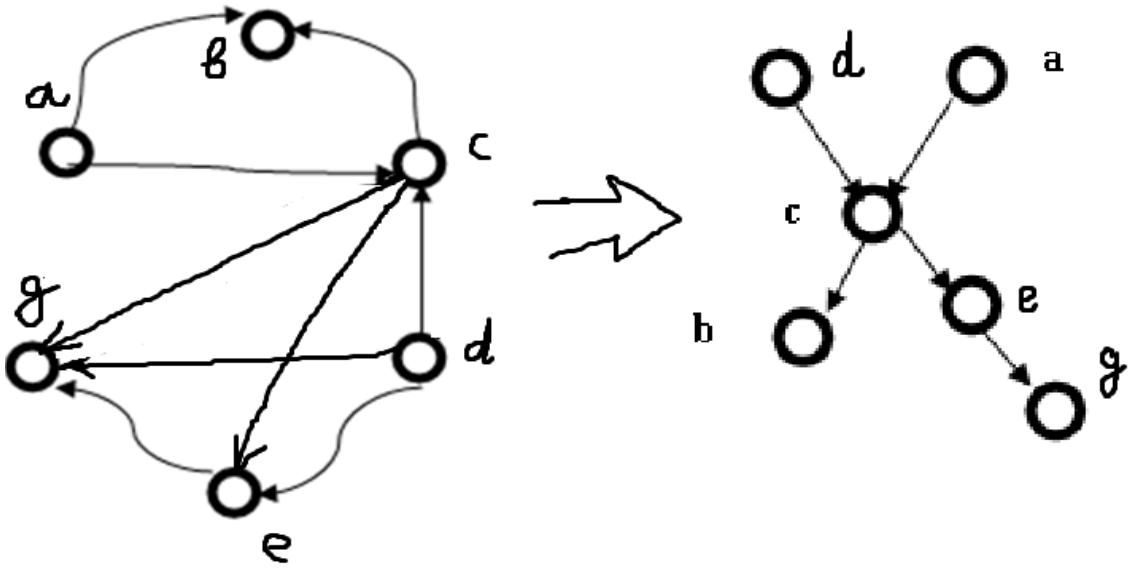
- Выделение связных подмножеств на множестве эталонов
- Линейной упорядочивание

связного подмножества

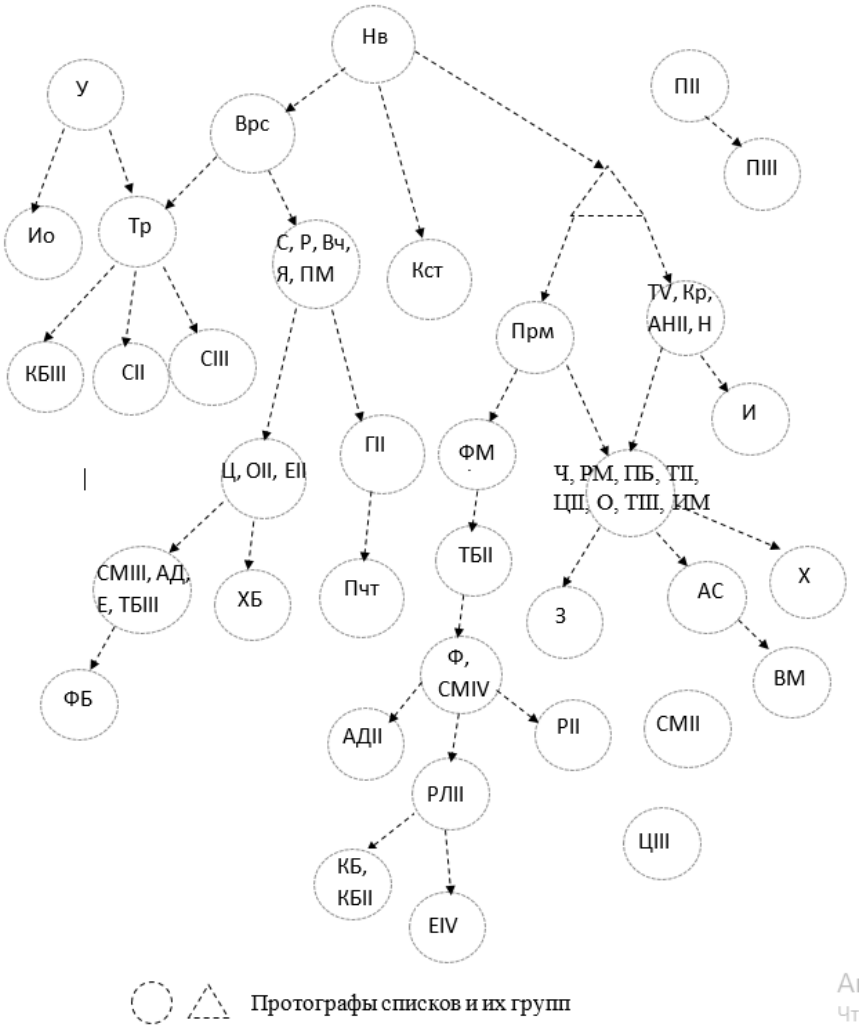


Хронологическое ранжирование текстов списков

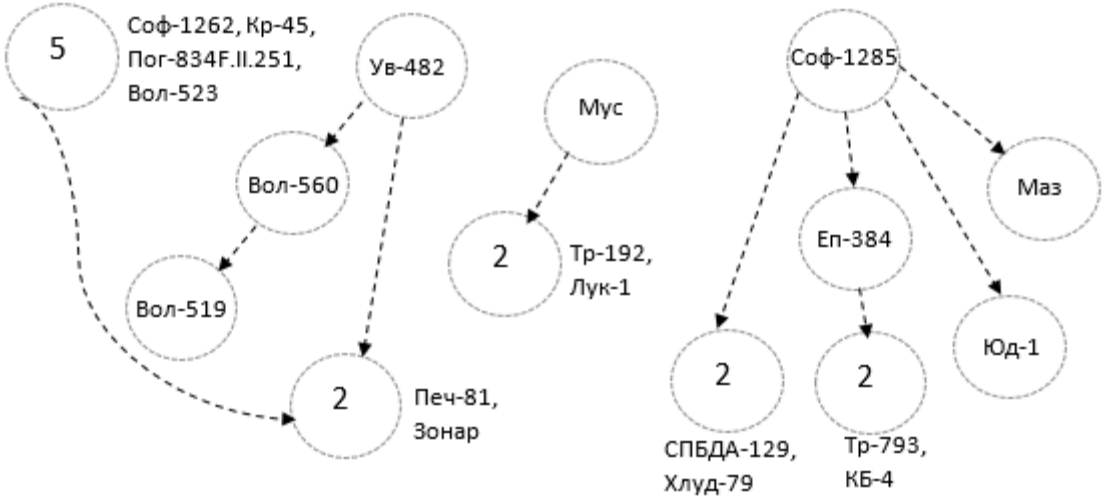
- Покрытие множества эталонов



Стемма ЗСЛ (порог 65, 84, 91%)



Стемма «Предсловия Покаянию» (порог 65%)



Проблемы и задачи

- Выбор универасального оптимального порога?
- Исследование устойчивости матрицы нечеткого отношения предпочтения
- Построение “дружественного” интерфейса блока сличения списков