OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822

ИЕРАРХИЯ НЕЧЕТКИХ ПОНЯТИЙ ДЛЯ ОПИСАНИЯ ИСТОРИЧЕСКОЙ ЛИЧНОСТИ

Глоба Л.С.[*], Островский С.А.[*], Попова М.А.[**], Стрижак А.Е.[**], Терновой М.Ю.[*]

[*] *Национальный технический университет Украины «Киевский политехнический интститут», г. Киев, Украина*

lgloba@its.kpi.ua

ternovoy@its.kpi.ua

ostazzz92@gmail.com

[**] *Институт телекоммуникаций и глобального информационного пространства НАН Украины, г. Киев, Украина*

sae953@gmail.com

pma1701@gmail.com

В работе представлен подход к построению иерархии нечетких понятий для описания исторической личности. Построение такой иерархии состоит из нескольких этапов, включающих формирование нечеткого контекста на основе литературных биографических источников, последующего применения анализа нечетких формальных понятий, кластерного анализа и затем создания иерархии понятий. Представлен фрагмент иерархии понятий, описывающей украинского поэта Шевченко Т.Г., полученной с использованием описанного подхода. Показан пример использования данного подхода в рамках тематического портала знаний.

Ключевые слова: иерархия понятий, нечеткий контекст, формальный анализ нечетких понятий.

Введение

Как правило, жизнь и деятельность известных исторических личностей описывается в различных биографических книгах, статьях и т.п. Эти источники могут, как пересекаться по содержанию, так и иметь свое особое видение данной личности. С другой стороны, специализированные порталы знаний должны иметь возможность представлять информацию о конкретной личности в структурированном виде [Borovikova, 2012], а также давать возможность описать историческую личность несколькими наиболее характерными ключевыми словами. Одним из возможных вариантов представления такого описания является представление в виде иерархии понятий [Ganter, 1999].

Отметим, что некоторые слова могут в большей степени описывать данную историческую личность, а другие – в меньшей степени. Такого рода неопределенность хорошо описывается при помощи математического аппарата нечетких множеств и нечеткой логики [Заде, 1976], [Новак, 2006]. В данной работе будет рассматриваться процесс построения описания исторической личности в виде иерархии нечетких понятий.

1. Постановка задачи

Анализ формальных понятий [Ganter, 1999] является методом анализа данных и представления знаний. В работе [Quan Thanh Tho, 2006] данный метод расширен за счет внедрения математического аппарата нечеткой логики и сформулирован анализ нечетких формальных понятий. Это позволило добавить возможность описания информации, обладающей неопределенностью. Анализ нечетких формальных понятий положен в основу описываемого в работе подхода.

Постановку задачи для построения описания исторической личности можно сформулировать следующим образом.

Дано: $D = \left\{ d_i \mid i = \overline{1,n} \right\}$ – множество книг, статей и т.п., описывающих историческую личность.

Найти: $(H, \angle)$ – иерархия понятий, где H - конечное множество понятий, $\angle$ - отношение частичного порядка на множестве H .

Решение данной задачи состоит в последовательном решении нескольких подзадач:

- построение нечеткого формального контекста исходя из множества документов D;
- определение множества нечетких понятий;
- построение решетки нечетких понятий;
- выделение кластеров в решетке нечетких понятий;
- формирование иерархии понятий.

Ниже рассмотрено последовательное решение этих задач на примере построения иерархии понятий (таксономий) жизнеописания для украинского поэта и общественного деятеля Шевченко Т.Г.

2. Построение нечеткого контекста для описания исторической личности

Под нечетким формальным контекстом будем понимать тройку $K=(G,M,I=\varphi(G\times M))$, где G - множество объектов, M - множество атрибутов, I - нечеткое множество на домене $G\times M$. Каждому отношению $(g,m)\in I$ соответствует значение функции принадлежности $\mu(g,m)\in[0;1]$.

В рассматриваемом случае множество объектов совпадает с множеством документов $G=D$. В качестве множества атрибутов будет выступать множество входящих в документы наиболее часто встречающихся слов $M=W=\{word_i\}$, таким образом, что всем словоформам ставится в соответствие одно слово. Тогда нечеткий формальный контекст будет записываться так $K=(D,W,I=\varphi(D\times W))$.

Решение задачи нахождения нечеткого формального контекста состоит из нескольких этапов. Вначале все документы проходят предварительную обработку, и каждому из них ставится в соответствие множество слов с количеством вхождения данного слова в данный документ. Таким образом, для каждого документа d_j мы получим $\{(word_{ji},number_{ji}),i=\overline{1,m_j}\}$, где $word_{ji}$ - слово, $number_{ji}$ - количество вхождения всех словоформ слова $word_{ji}$ в документ d_j, m_j - количество разных слов с учетом всех словоформ в документе d_j. Данная обработка текста проводилась с использованием системы КОНСПЕКТ [Величко, 2009].

Для построения нечеткого формального контекста для каждого документа выбирается m наиболее часто встречающихся слов. После этого проводится редактирование списка слов, таким образом, что каждому документу d_j вместо начального списка слов $WORD_j$ ставится в соответствие новое множество $WORD_j^*=\{word_{ji}\mid i=\overline{1,m_j^*}\}$, для которого выполняется условие:

$(\forall l\in L: word_{ji}\in WORD_l)\wedge(|L|\geq t)\wedge(j\in L)$, где

L - множество индексов $WORD_l$, в которые входит $word_{ji}$,

t - заданный порог, определяющий минимальное количество множеств, в которые входит $word_{ji}$.

Поскольку размер разных документов различен, то и количество наиболее употребляемых слов может отличаться в несколько раз. Этот факт необходимо учесть при определении значений функций принадлежности в нечетком формальном контексте. В данной работе предлагается определять значения функции принадлежности в нечетком формальном контексте следующим образом:

$$\mu(d_j,word_{ji})=\frac{\dfrac{number_{ji}}{\sum_i number_{ji}}}{\max\limits_{i,j}\dfrac{number_{ji}}{\sum_i number_{ji}}} \qquad (1)$$

При формировании нечеткого формального контекста в некоторых случаях имеет смысл задать порог на значение функций принадлежности и таким образом удалить не сильно влияющие связи.

Таким образом $W=\bigcup_j WORD_j^*$, $D=\bigcup_j d_j$, а отображение I определяется при помощи определенной в (1) функции принадлежности.

Нечеткий формальный контекст $K=(D,W,I)$, определенный для Т.Г. Шевченко показан в таблице 1. Имена строк соответствуют названиям документов: [Хлыпенко, 2008], [Рыльский, 1964], [Коваленко, 2000], [Меньшиков, 2008], [Бородин, 1984], обозначенным d1, d2, d3, d4, d5 соответственно. Имена столбцов соответствуют наиболее часто встречающимся словам. На пересечении стоит значение функции принадлежности, определяемой при помощи формулы (1).

Таблица 1 – Нечеткий формальный контекст

	Тарас	Кобзарь	Поэт	Стих	Человек
d1	0,72	0,5	0,35	0,35	0
d2	0	0	0,7	0,21	0,28
d3	0,52	0	1	0	0
d4	0	0,31	0	0	0,51
d5	0,35	0	0,89	0	0,29

3. Определение нечетких понятий и построение решетки

Для заданного нечеткого формального контекста $K=(D,W,I)$ и порога доверия T запишем несколько дополнительных понятий. Определим множества: $A^{*}=\{word\in W \mid \forall d\in A:\mu(d,word)\geq T\}$ для $A\subseteq D$ и $B^{*}=\{d\in D \mid \forall word\in B:\mu(d,word)\geq T\}$ для $B\subseteq W$.

Тогда нечеткое формальное понятие или просто нечеткое понятие для нечеткого формального контекста K и порога доверия T - это пара $(K_f=\varphi(A),B)$, где $A\subseteq D$, $B\subseteq W$, $A^{*}=B$ и $B^{*}=A$. Каждый документ $d\in\varphi(A)$ имеет функцию принадлежности, которая определяется формулой (2):

$$\mu_d=\min_{word\in B}\mu(d,word) \qquad (2)$$

Если $B=\varnothing$, то $\mu_d=1$ для всех документов. При таком определении формального понятия $(\varphi(A),B)$, множества A и B являются объемом и содержанием соответственно.

Таким образом, задача определения нечетких формальных понятий решается путем нахождения множеств A^{*} и B^{*}, проверки что $A^{*}=B$ и $B^{*}=A$, и последующего формирования нечетких понятий $(\varphi(A),B)$.

Для формирования решетки нечетких формальных понятий воспользуемся понятием похожести нечетких формальных понятий и алгоритмом построения их решетки, введенным в работе [Quan Thanh Tho, 2006].

Для определенного на предыдущих шагах нечеткого формального контекста и нечетких формальных понятий была получена решетка, которая использована для решения задачи описания исторической личности, показанная на рисунке 1.

4. Выделение кластеров в решетке и построение иерархии концептов

Под кластером понятий для решетки нечетких понятий L и порога подобия T_s будем понимать подрешетку S_L в L, которая имеет следующие свойства:

1. Наибольшее понятие C_S, которое входит в S_L, не совпадает ни с одним из своих суперпонятий.
2. Любое понятие $C\neq C_S$ в S_L должно иметь хотя бы одно суперпонятие $C'\in S_L$, такое, что мера похожести $E(C,C')>T_S$.

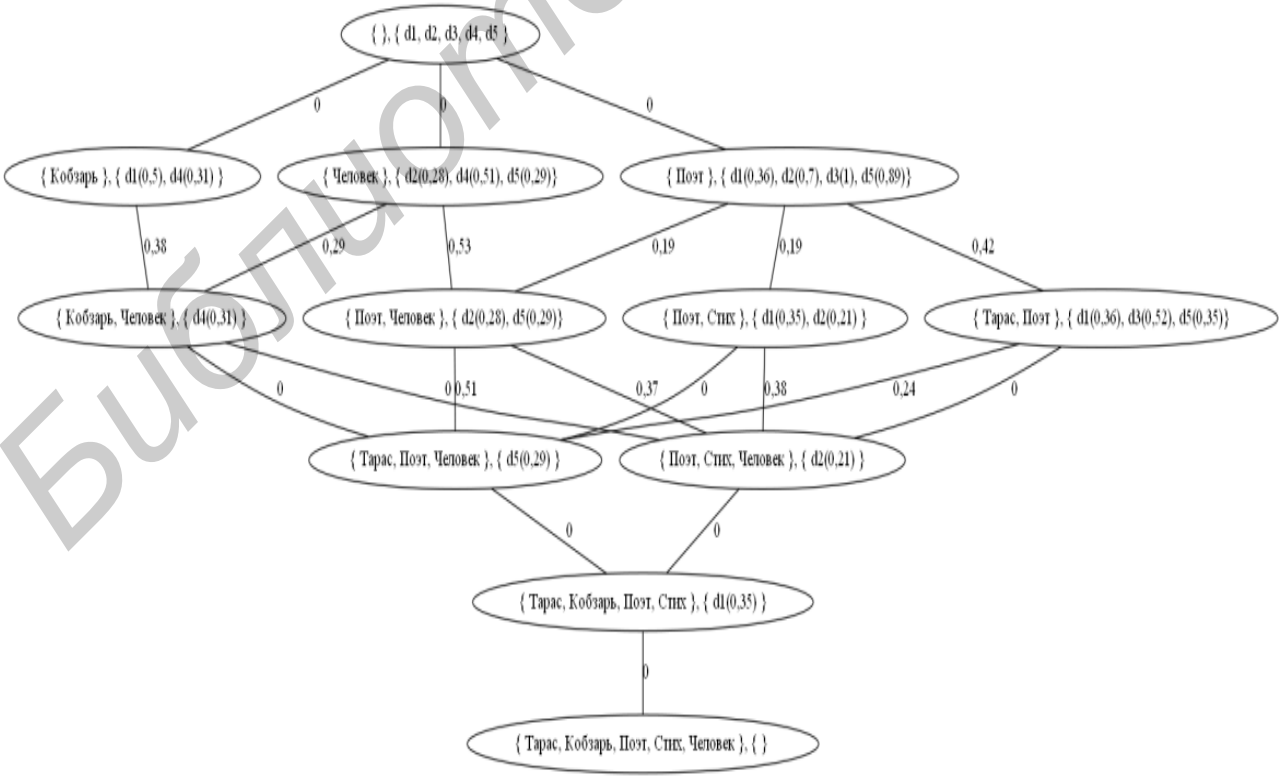

Рисунок 1 – Решетка нечетких формальных понятий

Под суперпонятием $C^{'}$ для конкретного понятия C здесь понимается более общее понятие, находящееся выше в решетке формальных понятий и связанное ребром с C.

Решетка с выделенными на ней кластерами показана на рисунке 2.

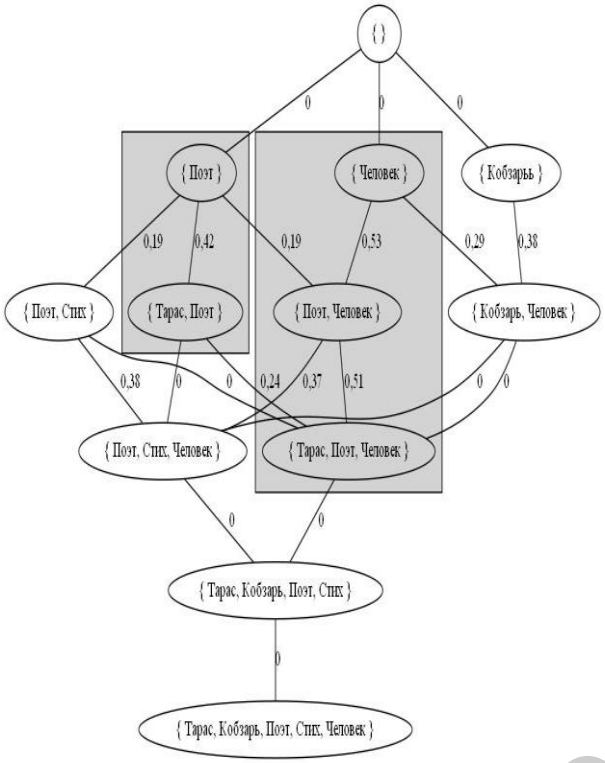

Рисунок 2 – Выделение кластеров в решетке нечетких формальных понятий

На основе полученной после кластеризации решетки нечетких формальных понятий формируется иерархия понятий, показанная на рисунке 3.

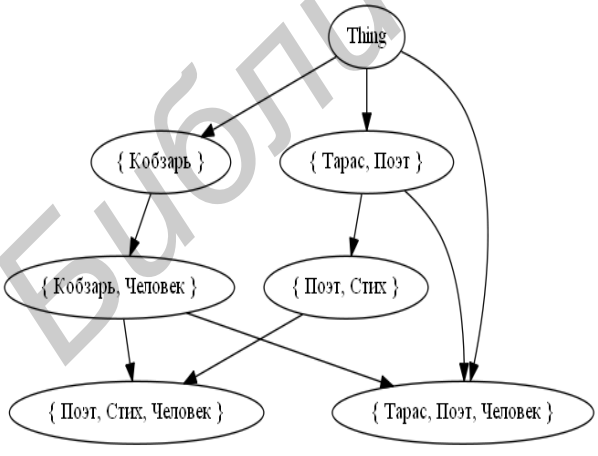

Рисунок 3 – Иерархия понятий

Полученная иерархия, по сути, является кратким описанием исторической личности, которое также несет дополнительную информацию о близости слов, используемых для описания.

5. Использование в порталах знаний

Рассмотрим расширение полученной выше таксономии для портала знаний, связанного с исследованием и анализом фактов жизни и деятельности Т. Г. Шевченко, с осмыслением и выделением главных факторов и причин тех или иных событий, связанных с биографией Кобзаря, а также их возможных последствий, влияние творчества на современников и будущие поколения.

Таксономии в портале знаний, обеспечивают группировку классов объектов предметной области. Таксономии формируются на основе установления отношений между понятиями и классами «часть - целое». Рассмотрим таксономии, вершины которых образуются понятиями, из которых в рамках ограничений решетки нечетких понятий можно сформулировать тавтологию [Мендельсон, 1971]. Для этого на основе определения кластера понятий для решетки нечетких понятий сформулируем следующее правило формирования тавтологий из понятий исследуемых предметных областей: <понятие S_L> включает в себя <понятие C_S>. Данное правило обеспечивает формирование таксономий.

Применим функцию принадлежности как образующую для формирования тавтологий из понятий таксономий. Тогда множество всех допустимых тавтологий состоит из утверждений о принадлежности понятий к определенной таксономии.

При обработке информационных источников на основе введенного правила были выделены понятия, объединѐнные в классы таксономий рассматриваемой предметной области, при чем из понятий сформированных таксономий могут быть сформулированы тавтологии в виде утверждений о событиях жизнедеятельности. Из множества всех таксономий выделены основные, которые приведенные ниже.

Тарас Григорьевич Шевченко (места пребывания в ипостаси):

- «Человек»;
- «Поэт»;
- «Писатель»;
- «Творец»;
- «Общественный/Политический деятель».

Маршруты путешествий:

- «Первое путешествие по Украине»;
- «Второе путешествие по Украине»;
- «Ссылка»;
- «Третье путешествие по Украине».

Почтение памяти:

- «Музеи»;
- «Памятники»;
- «Премии».

При этом отдельно взятое понятие, в свою очередь, может являться суперпонятием для других понятий. Например, понятие «Человек» является суперпонятием для понятий «Детство и юность», «Любовь Тараса Шевченко», «Смерть и похороны», а понятие «Писатель» - суперпонятие для понятий «Драматические произведения», «Повести», «Археологические заметки», «Записи народного творчества».

Для улучшения существующих таксономий на портале знаний предоставляется возможность дополнительного поиска по ключевым словам при помощи поисковых машин Интернет. Найденные таким образом электронные информационные источники могут быть обработаны аналогично для построения новых таксономий, которые в свою очередь, могут доопределить или расширить существующие таким образом, чтобы конструкция, начиная с понятия нижнего уровня и заканчивая понятием верхнего уровня таксономии, образовывала логическую цепочку.

Отображение понятий таксономии в портале знаний позволяет объединить понятия места и времени с понятиями фактов и событий в неизвестной до этого комбинации, под новым углом зрения.

Благодаря объединению различных типов баз данных в рамках портала знаний, в таксономии атрибуты объектов могут быть представлены не только в табличном виде, но и в текстовом, а также в виде тематических гиперссылок на распределенные в сети информационные ресурсы.

Заключение

Рассмотренный в работе подход к построению описаний исторических личностей позволяет учесть информацию из различных источников. Описания (иерархии нечетких понятий), полученные в результате, могут использоваться при построении и наполнении специализированных порталов знаний, а также формирования глобальной онтологии исторических личностей.

Данный подход не ограничен применением для описания исторических личностей, а может использоваться, например, для построения описания исторических событий, определения основного направления работы ученого и т.д. Основным условием является семантически схожая тематическая направленность исследуемых информационных ресурсов. Как видно из примера, именно тематическая направленность и влияет на значения функции принадлежности в нечетком формальном контексте, исследуемых распределенных информационных ресурсов.

Среди вопросов, которые необходимо дополнительно исследовать, можно выделить исследование зависимости построенных иерархий понятий от различных вариантов определения функции принадлежности в нечетком формальном контексте, а также развитие рассмотренного подхода для построения онтологий исторических личностей.

Библиографический список

[Бородин, 1984] Бородин, В. С. Т.Г. Шевченко. Биография (на украинском языке) / В. С. Бородин, Е. П. Кирилюк, В. Л. Смилянская, Е. С. Шаблиовский , В. Е. Шубравский // К., Наукова думка. – 588 с.

[Величко, 2009] Величко, В. Автоматизированное создание тезауруса терминов предметной области для локальных поисковых систем / В. Величко, П. Волошин, С. Свитла // “Knowledge – Dialogue – Solution” International Book Series “INFORMATION SCIENCE & COMPUTING”, Number 15. – FOI ITHEA Sofia, Bulgaria. - 2009. – pp.24-31.

[Заде, 1976] Заде, Л. Понятие лингвистической переменной и его применение к принятию приближенных решений / Л. Заде // М.: Мир, 1976. 166с.

[Коваленко, 2000] Коваленко, П. П. Сердце мое трудное, что у тебя болит?..(на украинском языке) / П.П. Коваленко // Черновцы, 2000. – 64с.

[Мендельсон, 1971] Мендельсон, Э. Введение в математическую логику /Э. Мендельсон // М. Наука, 1971. – 320 с.

[Меньшиков, 2008] Меньшиков, И.И. Поэтическое слово Кобзаря: Словарь лексических компонентов атрибутивных конструкций (на украинском языке) / И. И. Меньшиков, Н. В. Подмогильная // К., «Дирекция ФВД». – 212 с.

[Новак, 2006] Новак, В. Математические принципы нечёткой логики = Mathematical Principles of Fuzzy Logic. / В. Новак, И. Перфильева, И. Мочкрож // Физматлит, 2006. — 352 с.

[Рыльский, 1964] Рыльский, М. Тарас Шевченко. Биографический очерк (на украинском языке) / М. Ф. Рыльский, А. И. Дейч // К., Держлітвидав України. 1964. – 89 с.

[Хлыпенко, 2008] Тарас Шевченко в Кыргызстане: Научные, публицистические и художественные материалы на русском, киргизском и украинском языках / Ред-сост. Г.Н. Хлыпенко // Бишкек: Учкун, 2008. – 194 с.

[Borovikova, 2012] Borovikova, O. I. Methodology for knowledge portals development: background, foundations, experience of application, problems and prospects / O. I. Borovikova, L. S. Globa, R. L. Novogrudska, M.Y. Ternovoy, G. B. Zagorulko, Yu. A. Zagorulko // Bulletin of the Novosibirsk Computing Center, Issue: 34 (2012), pp. 73-92.

[Ganter, 1999] Ganter, B. Formal Concept Analysis: Mathematical Foundations / B. Ganter, R.Wille // Berlin-Heidelberg, Germany: Springer-Verlag, 1999.

[Quan Thanh Tho, 2006] Quan Thanh Tho. Automatic Fuzzy Ontology Generation for Semantic Web. / Quan Thanh Tho, Siu Cheung Hui, Tru Hoang Cao // IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 6, JUNE 2006.

FUZZY CONCEPT HIERARCHY DEVELOPMENT FOR HISTORICAL PERSONALITY DESCRIPTION

Globa L.S.[*], Ostrovskiy S.A.[*], Popova M.A.[**], Stryzhak O.Y.[**], Ternovoy M.Y.[*]

*National Technical University of Ukraine 'Kyiv Polytechnik Institute', Kyiv, Ukraine

lgloba@its.kpi.ua

ternovoy@its.kpi.ua

ostazzz92@gmail.com

**The Institute of Telecommunications and Global Information, National Academy of Science of Ukraine, Kyiv, Ukraine

sae953@gmail.com

pma1701@gmail.com

An approach for fuzzy concept hierarchy development for historical personality description is described in the paper. This approach consists of following steps: based on literary biographic sources determination of the fuzzy formal context, fuzzy formal concept analysis, fuzzy conceptual clustering and concept hierarchy generation. A part of generated with the using of the approach fuzzy concept hierarchy for Ukrainian poet T.G. Shevchenko are given. An example of the approach usage in knowledge portal development is described.

Introduction

Problem-oriented knowledge portals must have a possibility to present information about personality in structured manner [Borovikova, 2012] and give an opportunity to describe historic personality in a few key words.

One of the possible alternatives of such description is concept hierarchy [Ganter, 1999].

This paper describes an approach for historical personality description as fuzzy concept hierarchy.

Main Part

Formal concept analysis [Ganter, 1999] is a formal technique for data analysis and knowledge presentation. Quan Thanh Tho and others [Quan Thanh Tho, 2006] proposed a new technique called fuzzy formal concept analysis that combines fuzzy logic and formal concept analysis, in which the uncertainty information is directly represented by a real number of membership value in the range of [0,1]. Fuzzy formal concept analysis is the basis of the approach described in this paper.

Problem formalization for historical personality description follows.

Given:

$D = \{D_i \mid i = \overline{1,n}\}$ – a set of books, papers, another documents about historical personality.

Determine:

$(H, \angle)$ – concept hierarchy, where H - finite set of concepts, $\angle$ - is a partial order on H.

To solve this problem we need to solve several subproblems:

- to determine fuzzy formal context from the literary biographic sources D;
- to determine fuzzy formal concepts;
- to generate fuzzy formal concept lattice;
- to determine clusters in fuzzy formal concept lattice;
- to generate concept hierarchy.

Generated hierarchy is the brief description of the historical personality. It also has additional information about contextual similarity of words using for description.

Conclusion

Described approach for historical personality description allows one to take into account information from different sources. Obtained descriptions (fuzzy concept hierarchies) can be used in problem-oriented knowledge portal development. They also can be used for generation of global ontology for historic personality.

The application of this approach is not limited only for historical personality description. For example, it can be used for historical events description, for scientist's major research areas determination and so on. The main requirement for using the approach is the thematic similarity of the given information sources.