Кластеризация поисковых запросов Google Adwords для составления семантического ядра

Ахметшин Э.М., ассистент кафедры экономики и менеджмента, Казанский федеральный университет, Елабужский институт КФУ, г. Елабуга, Россия
Плотников А.В., канд. экон. наук, доцент кафедры «Менеджмент и маркетинг» Пермский национальный исследовательский политехнический университет, г. Пермь, Россия
Урасова А.А., канд. экон. наук, доцент, доцент кафедры государственного и муниципального управления, Пермский государственный национальный исследовательский университет, г. Пермь, Россия

Аннотация. Основная суть исследования состоит в том, чтобы поисковые запросы разбить на кластеры, которые содержат коммерческий (с высоким потенциалом для применения в интернет рекламе на поиске в Google и информационный (применимы для формирования семантического ядра и search engine optimization) интенты пользователей сети Интернет. Интенты позволят определить потенциал семантического ядра. В работе проведен кластерный анализ поисковых запросов по двум странам US и UK методом Ward. Несмотря на меньшее количество поисковых запросов на британском рынке, уровень конкуренции выше, чем в USA, следовательно, британский рынок контекстной рекламы оказался более агрессивным. В результате образовалось два кластера по cost per click и три кластера по Competition in pay per click.

Ключевые слова: контекстная реклама, интернет-маркетинг, Google Adwords, цифровая экономика, кластеризация, поисковый маркетинг, семантическое ядро.

Clustering of Google Adwords search queries for a semantic core development

Akhmetshin E.M., Kazan Federal University, Elabuga Institute of KFU,

Elabuga, Russia
Plotnikov A.V., Perm National Research Polytechnic University, Perm, Russia
Urasova A.A., Perm State National Research University, Perm, Russia

Annotation. The main essence of the study is to split the search queries into clusters that contain commercial (with high potential for use in Internet advertising on Google search) and informational (applicable to the formation of the semantic core and search engine optimization) intents of Internet users. A cluster analysis of search queries for two countries, US and UK, was conducted using the Ward method. Despite a smaller number of search queries in the UK market, the level of competition is higher than in the USA, which makes the UK content market more aggressive. As a result, two Cost per click clusters and three Competition in pay per click clusters were formed.

Keywords: contextual advertising, Google Adwords, digital economy, online marketing, clustering, search engine marketing, semantic core.

Введение

Все поисковые запросы пользователей имеют разнообразные задачи и выражаются в потребностях. Поиск и подбор поисковых запросов, подходящих для формирования семантического ядра, осуществляется за счет анализа потребительского поведения в поисковых системах, где в качестве объектов выступают запросы пользователей. Помимо этого, проводится статистический анализ запроса, выявляются и анализируются сайты конкурентов, проводится статистический сбор данных по интернет-ресурсу. По итогу анализа состав семантического ядра должен быть таким, чтобы полностью или максимально возможно соответствовать выбранной целевой аудитории и их конечным интересам.

Теоретические и практические аспекты по теме поисковых запросов были изучены в работах таких ученых как Jansen, B. J., Liu, Z., [1], Mehta, A. [2] & Goel G.[3].

Методы исследования

Кластерный анализ – метод классификационного анализа. Его цель – разделение массива изучаемых объектов, их свойств, характеристик на однородные совокупности – классы. Данный вид анализа относится к статистическим методам исследования, что обуславливает возможность изучать огромные массивы данных: большими по количеству могут быть как объекты, что исследуются, так и признаки, которые характеризуют такие объекты.

Преимущество метода в том, что он позволяет сгруппировать или разделить исследуемые переменные не по единому признаку, а по нескольким признакам. Помимо этого, кластерный анализ допускает исследовать разнообразные объекты, не ограничивая их по своей природе. Это отличает его от других математических и статистических методов, которые предоставляют возможность изучать ограниченное число переменных.

Каждый класс – это объединенные совокупности объектов, имеющих схожие свойства и характеристики. Задача метода – не только разделить имеющуюся совокупность переменных на m (m – целое) классов, а также соблюсти правило: каждый объект может принадлежать лишь к одному кластеру. Каждая переменная, попавшая в группу, должна иметь с другим объектом, входящим в группу, однородные признаки; объекты же, входящие в разные группы, должны иметь разнородные свойства, его характеризующие.

Если переменные, выбранные для исследования, разместить в пространстве с *n-измерением*, где *n* будет количеством признаков, которые характеризуют объекты, то схожесть или отличие исследуемых объектов или их признаков может быть описано расстоянием между точками. Исходя из этого, чем меньше расстояние между объектами в пространстве, тем более похожи эти объекты между собой; чем дальше расстояние – объекты отличаются один от одного.

Сейчас метод реализуется с помощью ЭВМ. Наиболее часто используемые алгоритмы провидения анализа – иерархические, или еще их называют древовидные. При этом есть агломеративные и итеративные дивизивные методы анализа.

Сущность иерархического агломеративного способа в том, что объекты объединяются по степени своей схожести друг с другом: сначала объединяются наиболее схожие, то есть расположенные более близко друг к другу, а потом уже удаленные друг от друга.

Иерархические же дивизивные алгоритмы позволяют осуществить кластерный анализ наоборот: сначала совокупность объектов объединяется по самому дальнему расстоянию, а потом уже по наиболее меньшему. За основу кластеризации почти все программы берут матрицу расстояния между объектами, то есть степень схожести исследуемых объектов.

Иерархические процедуры имеют недостатки, главный из которых – это достаточно сложный расчет и объемный массив данных, используемый для расчета. Поскольку на каждом этапе группировки машина должна рассчитывать матрицу расстояний, то иногда невозможно произвести группировку объектов. Особенно трудности возникают тогда, когда необходимо сгруппировать массив, содержащий более ста объектов.

Относительно того, как проводится агломеративный кластерный анализ, то его специфическая особенность в том, что изначально каждый исследуемый объект принимается за отдельный кластер. Потом уже каждый такой кластер объединяется с другим наиболее схожим кластером. На каждом этапе объединения рассчитывается новая матрица расстояний между объектами. Исследования может считаться завершенным в том случае, когда все объекты будут объединены, то есть сформируется один кластер. Программы, которые предлагают пользователю осуществить анализ, в основном предлагают результат в виде графического формата. Графическое изображение кластеризации называется дендрограммой.

В работе проведем анализ с использованием STATISTICA и реализуем агломеративными методами кластеризации. В метод входят древовидная кластеризация, метод k-средних.

Кластерный анализ с использованием метода k-средних

Особенность методологии – изначально выдвигается гипотеза, что исследуемая совокупность объектов разделяется в зависимости от признаков объектов или от переменных на m групп. Такая гипотеза может быть обоснована теоретическими или практическими данными, собственными догадками. В таком случае используемый метод *k-means* позволяет программе задать количество кластеров и получить желаемый результат. Количество кластеров можно изменять, получая разный результат.

Программа самостоятельно выбирает кластеры, а потом последовательно изменяет принадлежность различных объектов к данным кластерам. Программе необходимо добиться того, чтобы изменчивость объектов внутри кластеров была минимальна; а вот различия между самими кластерами были максимальны. Алгоритм определения принадлежности объекта к кластеру выглядит таким образом: система произвольно определяет центры кластеров, а потом вычисляет расстояние между данным центром каждого кластера и объектом. Объект, который имеет наименьшее расстояние к кластеру, к нему и приписывается. Когда все объекты распределены между кластерами, то система вычисляет среднее расстояние для каждого кластера. Таких показателей будет столько, сколько переменных было использовано для анализа, то есть *k*. Когда все *k* будут рассчитаны внутри кластера, то будет получен новый кластер. Потом алгоритм расчета повторяется: рассчитывается расстояние от каждого объекта до центра кластера, объект присоединяется к кластеру, с которым у него наименьшее расстояние. Программа проводит расчеты центров расстояний до тех пор, пока центры тяжести не перестанут менять координаты.

При провидении древовидного кластерного анализа можно использовать только категории для описания объектов. При методе k-средних для описания параметров объекта используют евклидово расстояние, что обуславливает необходимость стандартизации переменных перед началом анализа. Именно поэтому в методе k-средних переменные соизмеримы и представлены в единой интервальной шкале.

Условные обозначения

Рис. 1 – Рекламное объявление в Google Ads

The number of queries (Volume) – количество запросов в Google по USA/UK (чем частотнее (популярнее) запрос тем лучше для сайта). Высокочастотный запрос, который в рекламе стоит дорого, а в органической выдаче его легко продвинуть – это наиболее приоритетный запрос.

CPC (Cost per click) – стоимость 1 клика в долларах США, если пользователь введет запрос и по этому запросу покажется реклама и пользователь на нее кликнет, то с рекламодателя снимут указанную в этом столбце сумму. (Если клик дорогой, а «сложность фразы» от 0 до 20, то этому запросу или таким запросам важно уделять больше внимание и его надо скорее использовать для написания статей, чтобы сайт ранжировался по такому запросу и рекламодатель не платил за высокую стоимость клика.

Competition in PPC (Pay per click), %% – уровень конкуренции среди запросов. Чем больше сайтов используют данную фразу для показа рекламы, тем выше конкуренция.

Type of Ad - Ad placement type in relation to **search engine results page** (SERP):

1. Above;
2. Under;
3. Side.

Keyword Difficulty (Difficulty) – это уровень конкуренции этого запроса для попадания в топ10 органической выдачи. Чем выше уровень конкуренции (80-100) тем маловероятнее сайты смогут быть в выдаче топ10 Google по USA/UK. И

наоборот, если небольшие значения 0-20, то по этому запросу сайт можно легко продвинуть [4].

Results found – количество найденных страниц в результатах выдачи по запросу пользователей в поисковой системе Google.

Region_queries_count_last – queries, которые включают название региона.

Keyword length – N-gram, the number of words in query

Ad Text Length – Количество букв в тексте рекламного объявления.

Ad Title Length – Количество букв в Title рекламного объявления.

Ad Title Words – Количество слов в Title рекламного объявления.

Ad Text Words – Количество слов в тексте рекламного объявления.

Результаты исследования и их обсуждение

Проведем кластерный анализ и построим древовидные дендрограммы, используя программу STATISTICA.

При достижении задачи кластерного анализа – разбиение множества переменных на кластеры, необходимо соблюсти соответствие критериям оптимальности:

1. Каждый исследуемый объект должен входить в один класс;

2. Все переменные, которые принадлежат к кластеру, должны быть схожи;

3. Объекты, которые принадлежат к разным группам, должны быть не схожи.

Подобность объектов или их различие определяется с помощью m-мерного евклидового расстояния между переменными.

По поводу количества групп, гипотезу можно обосновать теоретическими наработками. Одновременно с этим, возможно определить количество кластеров и эмпирическим путем: разбивая совокупность объектов на m кластеры, где m равно от 1 до 5, сравнивая результаты и качество получаемых предложений. Исходя из этого, можно предположить, что показатели поисковых запросов попадут только в три группы таких запросов.

Алгоритм проведения анализа через программу выглядит следующим образом:

1. Определить центры будущих кластеров (случайно);

2. Между переменной и назначенным центром группы система определяет расстояние. Объект группируют с тем классом, с которым у него min расстояние;

3. Для полученной отдельной группы вычисляется среднее арифметическое. Их должно быть столько, сколько переменных выбрано для анализа объекта, то есть k штук. Рассчитанные средние становятся новыми координатами центра кластера;

4. Система снова проводит расчет расстояния от объекта до нового центра кластера. Объединяет объект к ближайшему кластеру.

Определяются центры групп. Расчеты проводятся до тех пор, пока центры не прекратят перемещение в плоскости или пространстве.

Вариант 1. Рассматривает поисковые запросы в Google и их CPC в US

Таблица 1

Дисперсионный анализ Варианта 1

Variable	Between SS	df	Within SS	df	F	Signif p
CPC	112765,2	1	52587,35	6478	13891,04	0,00

В таблице 1 приведены значения межгрупповых (Between SS) и внутригрупповых (Within SS) дисперсий признаков. Чем меньше значение внутригрупповой дисперсии и больше значение межгрупповой дисперсии, тем лучше признак характеризует принадлежность объектов к кластеру и тем «качественнее» наша кластеризация. Признаки с большими значениями *p* (например, больше 0,05) можно из процедуры кластеризации исключить. В нашем случае: для любого признака p<0,005, а значит никакой из рассматриваемых признаков исключать не будем.

Параметры F и p также характеризуют вклад признака в разделение объектов на группы. Лучшей кластеризации соответствуют большие значения первого и меньшие значения второго параметра. Из таблицы видно, что указанные выше наилучшие показатели соответствуют максимальной разнице (F – p). Рассмотрим средние значения для каждого кластера на линейном графике (рис. 2).

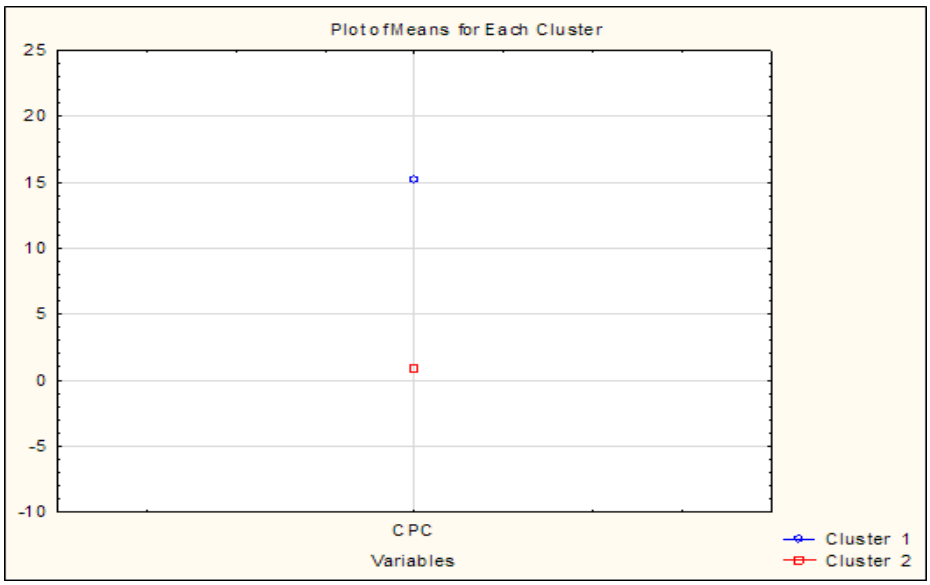

Рис. 2 – Средние значения двух кластеров (CPC в US)

Наглядно прослеживаются 2 кластера. Cluster 1 contains 5877 cases, cluster contains 603 cases, средние значения приведены в таблице 2.

Таблица 2

Средние значения по кластерам Вариант 1

Variable	Mean	Standard Deviation	Variance
CPC (Cluster 1)	15,21	7,38	54,52
CPC (Cluster 2)	0,85	1,83	3,36

Средние всех показателей значительно отличаются друг от друга. Это свидетельствует о качественном разбиении на группы. Как показывает график, расстояние между средними характеристик кластеров большое, также общее расстояние между центрами кластеров значительно, что свидетельствует об успешной кластеризации.

Целю кластерного анализа являлась деление на информационные и коммерческие (применяемые в рекламе) запросы. Полученные результаты полностью подтверждают проведенную классификацию:

Полученные результаты полностью подтверждают проведенную классификацию (три кластера):

– к кластеру под номером 1 относятся поисковые запросы, которые относятся к классу «реклама»,

– к кластеру под номером 2 относятся поисковые запросы, которые относятся к классу «информация»,

Проверим, действительно ли различны значения по переменным CPC US в выделенных кластерах. Для этого используем t-критерий Стьюдента.

Таблица 3

Т-критерий Стьюдента по кластерам CPC, US

Показатели методик	Ср. знач. Кластер 1	Ср. знач. Кластер 2	Т-критерий Стьюдента	Уровень значимости		Кол-во Кластер 1	Кол-во Кластер 2
				p			
The number of queries	115,75	77,60	0,85	0,40		603	5877
CPC	15,21	0,85	117,86	0,00	***	603	5877
Competition in PPC	62,88	23,95	34,65	0,00	***	603	5877
Keyword length	3,88	4,10	-3,59	0,00	***	603	5877
Type of Ad	1,77	2,26	-11,84	0,00	***	603	5877
Results found	198410208	114417902	3,02	0,00	**	603	5877
Region_queries_count_last	107,31	72,98	0,88	0,38		603	5877
Ad Text length	70,85	93,91	-0,59	0,55		603	5877
Ad Text Words	11,21	11,48	-0,22	0,82		603	5877
Ad Title length	121,30	130,58	-0,14	0,89		603	5877
Ad Title Words	11,86	12,52	-0,18	0,86		603	5877

Согласно Т-критерий Стьюдента два выделенных кластера существенно отличаются друг от друга по значениям переменных CPC, Competition in PPC, Keyword length, Type of Ad, Results found. И значимо не различны по значениям переменных The number of queries, Region_queries_count_last, Ad Text Words, Ad Text length, Ad Title Words, Ad Title Length. Следовательно, количество запросов не отражает реальную CPC и Competition in PPC.

Вариант 2. Рассматривает сложность продвижения поисковых запросов в контекстной рекламе Google в US

Таблица 4

Дисперсионный анализ Варианта 2

Variable	Between SS	df	Within SS	df	F	Signif p
Competition in PPC	4740859	2	561706,40	6477	27333,27	0,00

Для любого признака p<0,005, а значит никакой из рассматриваемых признаков исключать не будем.

Рассмотрим средние значения для каждого кластера на линейном графике (рис. 3).

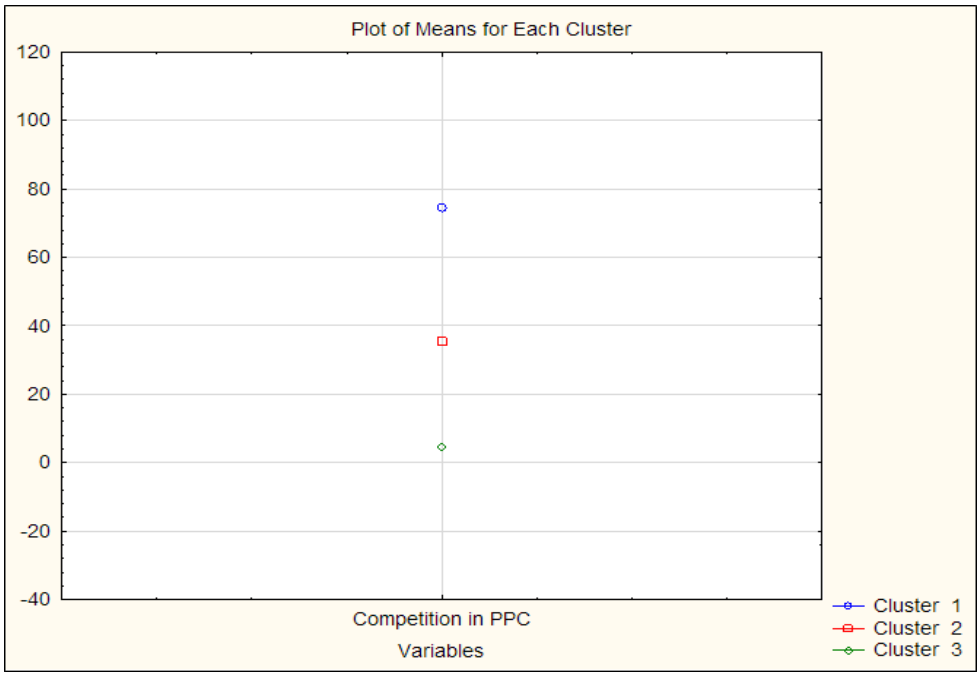

Рис. 3 – Средние значения трех кластеров (Competition in PPC в US)

Наглядно прослеживаются 3 кластера. Cluster 1 contains 1311 cases, cluster 2 contains 1860 cases, cluster 3 contains 3309 cases, средние значения приведены в таблице 4.

Таблица 5

Средние значения по кластерам Вариант 2

Variable	Mean	Standard Deviation	Variance
Competition in PPC (Cluster 1)	74,56	14,01	196,27
Competition in PPC (Cluster 2)	35,21	9,84	96,78
Competition in PPC (Cluster 3)	4,66	6,14	37,69

Средние всех показателей значительно отличаются друг от друга. Это свидетельствует о качественном разбиении на группы. Как показывает график, расстояние между средними характеристик кластеров большое, также общее расстояние между центрами кластеров значительно, что свидетельствует об успешной кластеризации.

Полученные результаты полностью подтверждают проведенную классификацию (три кластера):

– к кластеру под номером 1 относятся поисковые запросы, которые относятся к классу «реклама»,

– к кластеру под номером 2 относятся поисковые запросы, которые относятся к классу «реклама-информация»,

– к кластеру под номером 3 относятся поисковые запросы, которые относятся к классу «информация»,

Проверим, действительно ли различны значения по переменным PPC US в выделенных кластерах. Для этого используем t-критерий Стьюдента.

Таблица 6

Т-критерий Стьюдента по кластерам 1 и 2 PPC US

Показатели методик	**Ср. знач. Кластер 1**	**Ср. знач. Кластер 2**	**Т-критерий Стьюдента**	**Уровень значимости**		**Кол-во Кластер 1**	**Кол-во Кластер 2**
				p			
The number of queries	77,61	109,07	-0,69	0,49		1311	1860
CPC	6,28	2,67	15,95	0,00	***	1311	1860
Competition in PPC	74,56	35,21	92,94	0,00	***	1311	1860
Keyword length	4,22	4,11	2,25	0,02	*	1311	1860
Type of Ad	1,89	2,01	-3,40	0,00	***	1311	1860
Results found	99102788	125758609	-1,39	0,16		1311	1860
Region_queries_count_last	72,10	98,18	-0,69	0,49		1311	1860
Ad Text Length	124,52	83,40	1,13	0,26		1311	1860
Ad Text Words	12,65	11,41	1,15	0,25		1311	1860
Ad Title Length	127,34	196,93	-1,00	0,32		1311	1860
Ad Title Words	13,21	15,90	-0,64	0,52		1311	1860

Согласно Т-критерий Стьюдента два выделенных кластера существенно отличаются друг от друга по значениям переменных CPC, Competition in PPC, Keyword length, Type of Ad. И значимо не различны по значениям переменных The number of queries, Results found, Region_queries_count_last, Ad Text Words, Ad Text length, Ad Title Words, Ad Title Length. Следовательно количество запросов не отражает реальную CPC и Competition in PPC.

Таблица 7

Т-критерий Стьюдента по кластерам 1 и 3 PPC US

Показатели методик	Ср. знач. Кластер 1	Ср. знач. Кластер 3	Т-критерий Стьюдента	Уровень значимости		Кол-во Кластер 1	Кол-во Кластер 3
				p			
The number of queries	77,61	66,87	0,48	0,63		1311	3309
CPC	6,28	0,29	41,91	0,00	***	1311	3309
Competition in PPC	74,56	4,66	235,59	0,00	***	1311	3309
Keyword length	4,22	4,01	4,52	0,00	***	1311	3309
Type of Ad	1,89	2,46	-19,28	0,00	***	1311	3309
Results found	99102788	129416922	-1,36	0,17		1311	3309
Region_queries_count_last	72,10	65,42	0,31	0,76		1311	3309
Ad Text Length	124,52	83,50	1,30	0,19		1311	3309
Ad Text Words	12,65	11,00	1,74	0,08		1311	3309
Ad Title Length	127,34	92,87	0,82	0,41		1311	3309
Ad Title Words	13,21	10,22	1,28	0,20		1311	3309

Согласно Т-критерий Стьюдента два выделенных кластера существенно отличаются друг от друга по значениям переменных CPC, Competition in PPC, Keyword length, Type of Ad. И значимо не различны по значениям переменных The number of queries, Results found , Region_queries_count_last, Ad Text Words, Ad Text length, Ad Title Words, Ad Title Length. Следовательно, количество запросов не отражает реальную CPC и Competition in PPC.

Таблица 8

Т-критерий Стьюдента по кластерам 2 и 3 PPC US

Показатели методик	Ср. знач. Кластер 2	Ср. знач. Кластер 3	Т-критерий Стьюдента	Уровень значимости		Кол-во Кластер 2	Кол-во Кластер 3
				p			
The number of queries	109,07	66,87	1,25	0,21		1860	3309
CPC	2,67	0,29	26,67	0,00	***	1860	3309
Competition in PPC	35,21	4,66	137,29	0,00	***	1860	3309
Keyword length	4,11	4,01	2,56	0,01	*	1860	3309
Type of Ad	2,01	2,46	-16,92	0,00	***	1860	3309
Results found	125758609	129416922	-0,18	0,85		1860	3309
Region_queries_count_last	98,18	65,42	1,11	0,27		1860	3309
Ad Text Length	83,40	83,50	0,00	1,00		1860	3309
Ad Text Words	11,41	11,00	0,54	0,59		1860	3309
Ad Title Length	196,93	92,87	2,25	0,02	*	1860	3309
Ad Title Words	15,90	10,22	2,43	0,02	*	1860	3309

Согласно Т-критерий Стьюдента два выделенных кластера существенно отличаются друг от друга по значениям переменных CPC, Competition in PPC,

Keyword length, Type of Ad, Ad Title Words, Ad Title Length. И значимо не различны по значениям переменных The number of queries, Results found, Region_queries_count_last, Ad Text Words, Ad Text length. Следовательно, количество запросов не отражает реальную CPC и Competition in PPC.

Вариант 3. Рассматривает поисковые запросы в Google и их CPC в UK

Таблица 9

Дисперсионный анализ Варианта 3

Variable	Between SS	df	Within SS	df	F	Signif p
CPC	4740859	2	561706,40	6477	27333,27	0,00

Для любого признака p<0,005, а значит никакой из рассматриваемых признаков исключать не будем. Значения межгрупповых (Between SS) и внутригрупповых (Within SS) дисперсий признаков переменной CPC больше в 3 варианте, чем в варианте 1. Что характеризует кластерный анализ переменных для страны US в варианте 1 как более точный, чем кластерный анализ варианта 3 для страны UK.

Рассмотрим средние значения для каждого кластера на линейном графике (рис. 4).

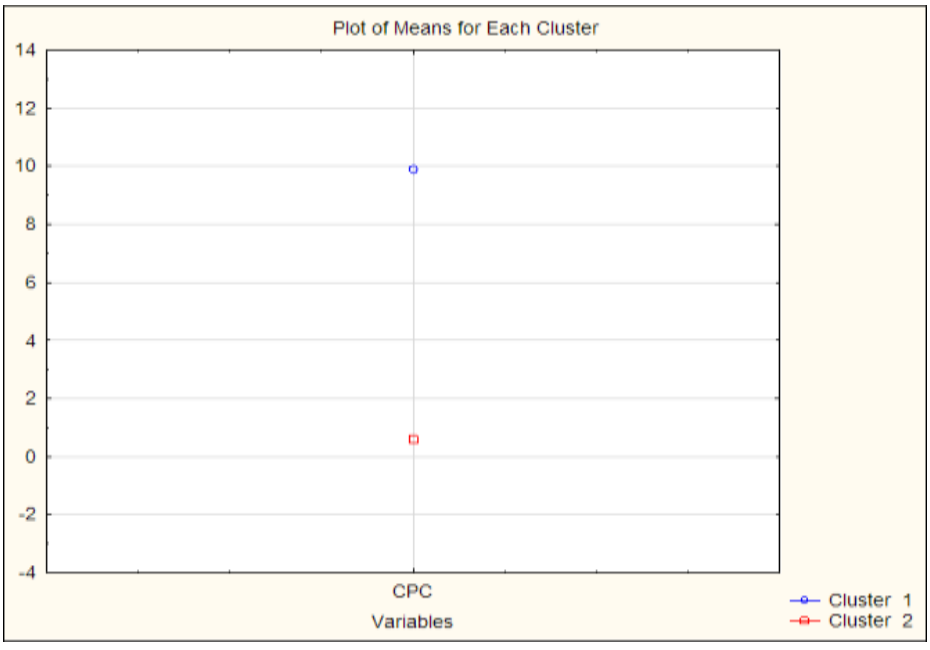

Рис. 4 – Средние значения трех кластеров (CPC в US)

На рисунке можно выделить 2 кластера. Cluster 1 contains 175 cases, cluster contains 1477 cases, средние значения приведены в таблице 6.

Таблица 10

Средние значения по кластерам Вариант 3

Variable	Mean	Standard Deviation	Variance
CPC (Cluster 1)	9,89	5,66	32,01
CPC (Cluster 2)	0,59	1,23	1,52

Значения кластеров в варианте 3 незначительно ниже, чем аналогичные показатели в варианте 1. Это значит, запросы в поисковой системе Google можно разделить на 1) запросы с коммерческим потенциалом, которые можно использовать в настройке контекстной рекламы на поиске; 2) запросы с информативным потенциалом для search engine optimization and Display Ads.

Проверим, действительно ли различны значения по переменным CPC US в выделенных кластерах. Для этого используем t-критерий Стьюдента.

Таблица 11

Т-критерий Стьюдента по кластерам CPC UK

Показатели методик	**Ср. знач. Кластер 1**	**Ср. знач. Кластер 2**	**Т-критерий Стьюдента**	**Уровень значимости**		**Кол-во Кластер 1**	**Кол-во Кластер 2**
				p			
The number of queries	92,74	111,64	-0,19	0,85		175	1477
CPC	9,89	0,59	53,42	0,00	***	175	1477
Competition in PPC	77,28	28,62	19,59	0,00	***	175	1477
Keyword length	3,55	3,81	-2,62	0,01	**	175	1477
Type of Ad	1,98	2,01	-0,31	0,76		175	1477
Results found	301778900	152304137	2,65	0,01	**	175	1477
Region_queries_count_last	85,37	104,43	-0,19	0,85		175	1477
Ad Text Length	76,46	76,95	-0,12	0,91		175	1477
Ad Text Words	12,18	12,10	0,13	0,90		175	1477
Ad Title Length	51,63	93,26	-0,46	0,65		175	1477
Ad Title Words	8,73	9,55	-0,34	0,73		175	1477

Согласно Т-критерий Стьюдента два выделенных кластера существенно отличаются друг от друга по значениям переменных CPC, Competition in PPC, Keyword length, Results found. И значимо не различны по значениям переменных The number of queries, Type of Ad, Region_queries_count_last, Ad Text Words, Ad Text length, Ad Title Words, Ad Title Length. Следовательно, количество запросов не отражает реальную CPC и Competition in PPC.

Вариант 4. Рассматривает сложность продвижения поисковых запросов в контекстной рекламе Google в UK

Таблица 12

Дисперсионный анализ Варианта 4

Variable	Between SS	df	Within SS	df	F	Signif p
Competition in PPC	1810827	2	152761,2	1649	9773,599	0,00

Для любого признака p<0,005, а значит никакой из рассматриваемых признаков исключать не будем. Значения межгрупповых (Between SS) и внутригрупповых (Within SS) дисперсий признаков переменной CPC больше в 4 варианте, чем в варианте 2. Что характеризует кластерный анализ переменных для страны US в варианте 2 как более точный, чем кластерный анализ варианта 4 для страны UK. Рассмотрим средние значения для каждого кластера на линейном графике (рис. 5).

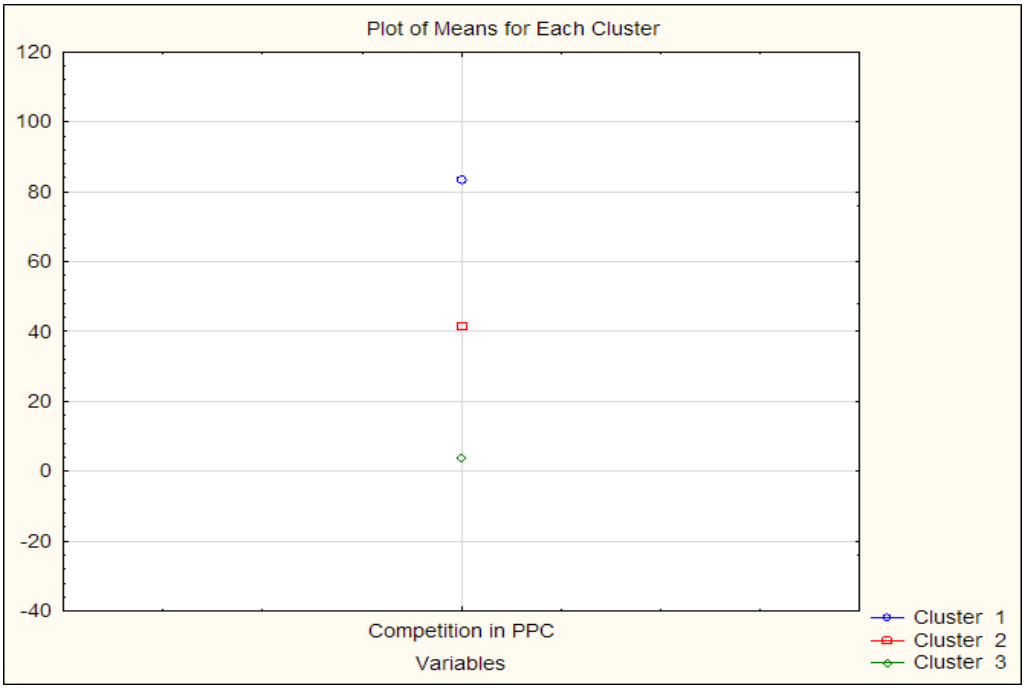

Рис. 5 – Средние значения трех кластеров (Competition in PPC, UK)

На рисунке можно выделить 3 кластера. Cluster 1 contains 426 cases, cluster 2 contains 404 cases, cluster 3 contains 822 cases, средние значения приведены в таблице 8.

Таблица 13

Средние значения по кластерам Вариант 4

Variable	Mean	Standard Deviation	Variance
Competition in PPC (Cluster 1)	83,64	12,06	145,35
Competition in PPC (Cluster 2)	41,69	11,73	137,69
Competition in PPC (Cluster 3)	4,05	6,58	43,24

Значения кластеров в варианте 2 незначительно ниже, чем аналогичные показатели в варианте 2 (USA). Это значит, что несмотря на меньшее количество поисковых запросов на британском рынке, уровень конкуренции выше, чем в USA, следовательно, британский рынок контекстной рекламы более агрессивен.

Проверим, действительно ли различны значения по переменным PPC US в выделенных кластерах. Для этого используем t-критерий Стьюдента.

Таблица 14

Т-критерий Стьюдента по кластерам 1 и 2 Competition in PPC, UK

Показатели методик	Ср. знач. Кластер 1	Ср. знач. Кластер 2	Т-критерий Стьюдента	Уровень значимости		Кол-во Кластер 1	Кол-во Кластер 2
				p			
The number of queries	77,04	61,24	1,90	0,06		426	404
CPC	4,30	1,57	8,96	0,00	***	426	404
Competition in PPC	83,64	41,69	50,75	0,00	***	426	404
Keyword length	3,74	3,69	0,60	0,55		426	404
Type of Ad	2,00	1,73	3,91	0,00	***	426	404
Results found	215979649	161691366	0,93	0,35		426	404
Region_queries_count_last	74,95	55,99	2,18	0,03	*	426	404
Ad Text Length	71,68	72,30	-0,19	0,85		426	404
Ad Text Words	11,47	11,51	-0,09	0,93		426	404
Ad Title Length	205,59	50,26	1,39	0,16		426	404
Ad Title Words	12,77	8,37	1,50	0,13		426	404

Согласно Т-критерий Стьюдента два выделенных кластера существенно отличаются друг от друга по значениям переменных CPC, Competition in PPC, Type of Ad, Region_queries_count_last. И значимо не различны по значениям переменных The number of queries, Keyword length, Results found, Ad Text Words, Ad Text length, Ad Title Words, Ad Title Length. Следовательно, количество запросов не отражает реальную CPC и Competition in PPC.

Таблица 15

Т-критерий Стьюдента по кластерам 1 и 3 Competition in PPC, UK

Показатели методик	Ср. знач. Кластер 1	Ср. знач. Кластер 3	Т-критерий Стьюдента	Уровень значимости		Кол-во Кластер 1	Кол-во Кластер 3
				p			
The number of queries	77,04	150,32	-0,86	0,39		426	822
CPC	4,30	0,17	21,15	0,00	***	426	822
Competition in PPC	83,64	4,05	150,89	0,00	***	426	822
Keyword length	3,74	3,84	-1,36	0,17		426	822
Type of Ad	2,00	2,14	-2,38	0,02	*	426	822
Results found	215979649	146513230	1,78	0,08		426	822
Region_queries_count_last	74,95	139,46	-0,76	0,45		426	822
Ad Text Length	71,68	81,85	-3,15	0,00	**	426	822
Ad Text Words	11,47	12,74	-2,77	0,01	**	426	822
Ad Title Length	205,59	47,32	2,03	0,04	*	426	822
Ad Title Words	12,77	8,29	2,17	0,03	*	426	822

Согласно Т-критерий Стьюдента два выделенных кластера существенно отличаются друг от друга по значениям переменных CPC, Competition in PPC, Type of Ad, Ad Text Words, Ad Text length, Ad Title Words, Ad Title Length. И значимо не различны по значениям переменных The number of queries, Keyword length, Results found, Region_queries_count_last, Ad Text Words, Ad Text length, Ad Title Words, Ad Title Length. Следовательно, количество запросов не отражает реальную CPC и Competition in PPC.

Таблица 16

Т-критерий Стьюдента по кластерам 2 и 3 Competition in PPC, UK

Показатели методик	Ср. знач. Кластер 2	Ср. знач. Кластер 3	Т-критерий Стьюдента	Уровень значимости		Кол-во Кластер 2	Кол-во Кластер 3
				p			
The number of queries	61,00	150,00	-1,02	0,31		404	822
CPC	2,00	0,00	13,08	0,00	***	404	822
Competition in PPC	42,00	4,00	71,86	0,00	***	404	822
Keyword length	4,00	4,00	-1,90	0,06		404	822
Type of Ad	2,00	2,00	-6,84	0,00	***	404	822
Results found	161691366	146513230	0,38	0,70		404	822
Region_queries_count_last	56,00	139,00	-0,96	0,34		404	822
Ad Text Length	72,00	82,00	-3,07	0,00	**	404	822
Ad Text Words	12,00	13,00	-2,78	0,01	**	404	822
Ad Title Length	50,00	47,00	3,35	0,00	***	404	822
Ad Title Words	8,00	8,00	0,48	0,63		404	822

Согласно Т-критерий Стьюдента два выделенных кластера существенно отличаются друг от друга по значениям переменных CPC, Competition in PPC, Type of Ad, Ad Text Words, Ad Text length, Ad Title Words. И значимо не различны по значениям переменных The number of queries, Keyword length, Results found, Region_queries_count_last, Ad Title Length. Следовательно, количество запросов не отражает реальную CPC и Competition in PPC.

Мы убедились в том, что данные о переменных качественно определяют принадлежность к классам и пригодны для дальнейшего исследования. Кластерный анализ позволяет провести объективную классификацию поисковых запросов, которые охарактеризованы рядом признаков (The number of queries, CPC, Competition in PPC, Keyword length). Из этого можно сделать следующую рекомендацию: третий кластер можно не использовать в качестве ключевых слов для запуска рекламных кампаний на поиске.

Каждый метод имеет недостатки использования. Кластерный анализ не исключение. Минусы метода:

- Количество классов и их состав обуславливается теми параметрами, которые описывают совокупность;
- При стандартизации данных и при других методах сведения данных к более компактному формату, может страдать качество переменных;
- Некоторые свойства объектов, попавших в определенный кластер, могут искажаться или терять свою значимость за счет замены их свойствами целого кластера.

Изучим результаты кластеризации поисковых запросов, которые показали достаточно неоднозначные результаты. Цель кластеризации по древовидному типу в том, чтобы исследуемые объекты объединить в большие по объемы классы. Для объединения используем расстояние между такими объектами, входящими в кластер. По итогу анализа пользователь должен получить иерархическое дерево.

Более подробно изучим древовидную диаграмму. По построенному дереву можно увидеть, что дерево начинает рост с каждого исследуемого значения

объекта. Поэтапно требования к схожести классов становятся более «мягкими», что позволяет объединять в единый блок менее схожие переменные. С каждым шагом порог относительно требований, выдвигаемых к единой группе, снижается.

По итогу упрощения порога все большее количество изучаемых объектов объединяется в меньшее количество кластеров. На завершающем этапе образуется единый кластер.

Представленные диаграммы кластерного анализа имеют горизонтальные оси, которые представляют информацию о расстоянии между объектами. Например, на древовидных дендрограммах расстояние между объектами и кластерами представляется на вертикальных осях.

Если представленные для анализа объекты, их параметры, свойства и т.д. изначально имеют четко выраженную структуру, то тогда и при построении иерархического дерева такая структура будет изображена четко выраженными ветвями. По итогу такого анализа можно выявить кластеры, а потом описать их свойства, интерпретируя полученные данные.

Меры расстояния

Когда необходимо выявить кластеры не схожих объектов или определить расстояние между ними, то целесообразно использовать метод древовидной кластеризации. Расстояния между объектами при этом могут определяться как в одномерном, так и многомерном пространстве.

Для этого нужно соблюдать правила объединения и связи.

Метод Ward. Суть метода в том, что он использует для расчета расстояний между группами дисперсионный анализ. Это отличает его от других методов кластерного анализа. При использовании метода Ward минимизируется сумма квадратов по двум кластерам. Во время кластеризации объединяются те группы, которые приводят к минимальному росту функции, то есть к минимальному росту суммы квадратов отклонения внутри кластера. [5] [6] [7] Цель метода – объединить в единую группу объекты, которые расположены ближе всего к друг другу. И несмотря на эффективность метода, его недостаток в том, что он создает

небольшие группы. И теперь, используя данный метод, а также данные таблиц изучим горизонтальную древовидную диаграмму по каждому варианту.

По каждому изученному варианту представлены результаты кластеризации. Результаты в Приложении 1-4. По каждому из полученных кластеров представлен дополнительно Plot of linkage distances across steps. Графическое представление результатов помогает выявить кластеры и их количество (на графике есть как точка объединения, так и номер шага k, где такое объединение произошло). По итогу графика получаем, что количество кластеров равно n-k, где n – массив данных, количество объектов для построения анализа.

Чтобы определить кластеры с учетом изучаемых переменных обязательно нужно определить предельное расстояние между объектами. После того, как такое предельное расстояние (порог) определено, то необходимо провести перпендикулярную прямую через точку, определяющую расстояние. По итогу построения перпендикуляра подсчитывается количество точек пересечения. Сколько точек пересечения перпендикуляра и ветвей будет выявлено, столько будет и кластеров. Объекты же, оказавшиеся на данной ветви отсечения, будут теми объектами, которые входят в кластер.

Заключение

Благодаря анализу произведено деление поисковых запросов на кластеры с низким и высоким CPC. Наглядно видно какие фразы относятся к кластеру «реклама» для US, а какие для UK. Так же слова разбиты на кластеры по уровню сложности. Различается высокая, средняя и низкая сложность слов для каждой из страны, что позволяет определять поисковые запросы, которые приоритетнее использовать в рекламе, а по каким продвигаться в SEO.

Исследование выполнено при финансовой поддержке РФФИ и Пермского края в рамках научного проекта № 18-410-590007.

Библиографический список

1. Jansen, B.J., Liu, Z., Weaver, C., Campbell, G., & Gregg, M. (2011). Real time search on the web: Queries, topics, and economic value. Information Processing & Management, 47(4), 491-506.

2. Mehta, A. (2012). Online matching and Adwords.

3. Goel, G., & Mehta, A. (2008, January). Online budgeted matching in random input models with applications to adwords. In Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 982-991). Society for Industrial and Applied Mathematics.

4. Serpstat [URL] www.serpstat.com

5. Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301), 236-244.

6. Hervada-Sala, C., & Jarauta-Bragulat, E. (2004). A program to perform Ward's clustering method on several regionalized variables. Computers & Geosciences, 30(8), 881-886.

7. Batagelj, V. (1988). Generalized Ward and related clustering problems. Classification and related methods of data analysis, 67-74.

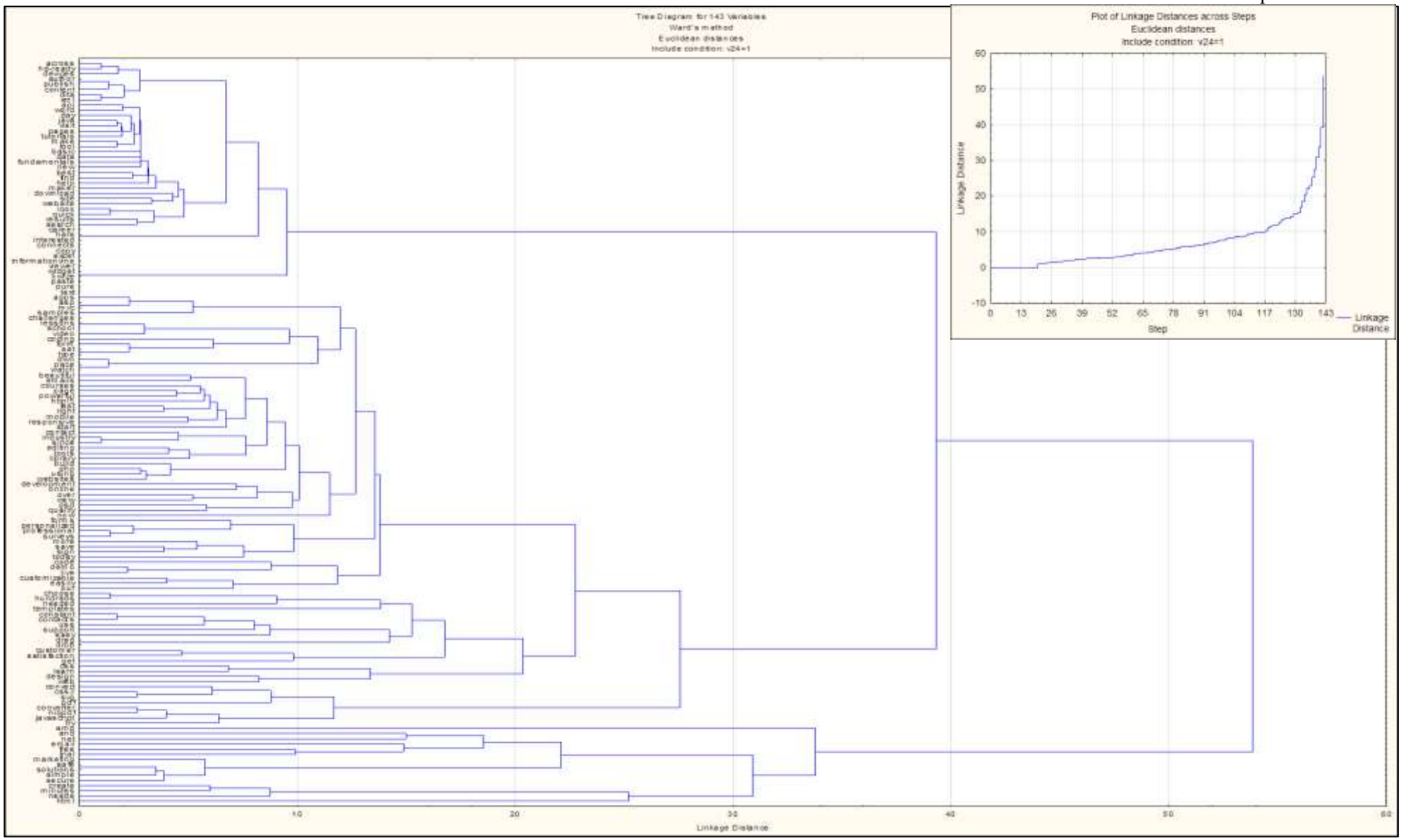

Рис. 1 – Вариант 1. Кластерный анализ поисковых запросов по переменной CPC. Страна US

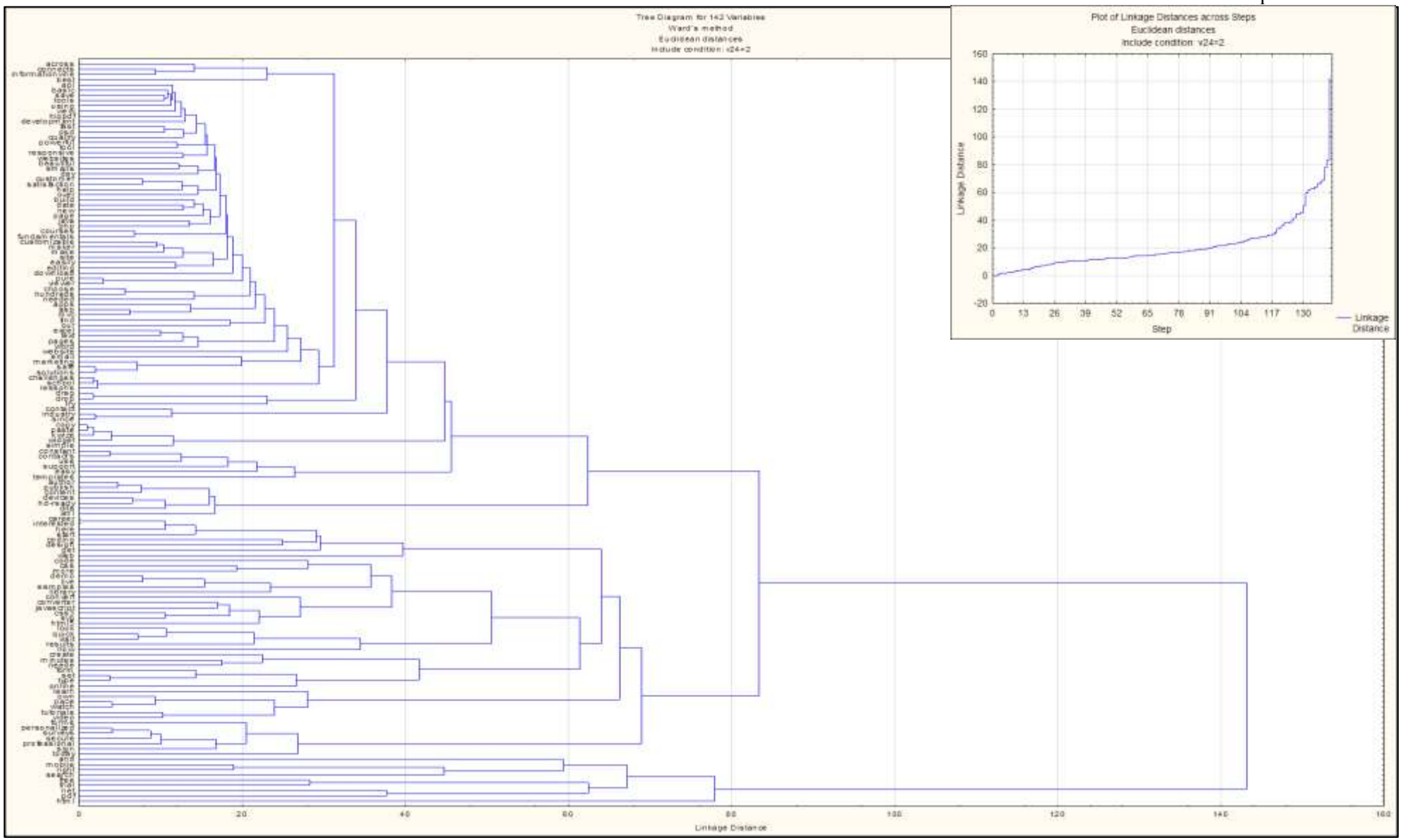

Рис. 2 – Вариант 1. Кластерный анализ поисковых запросов по переменной CPC. Страна US

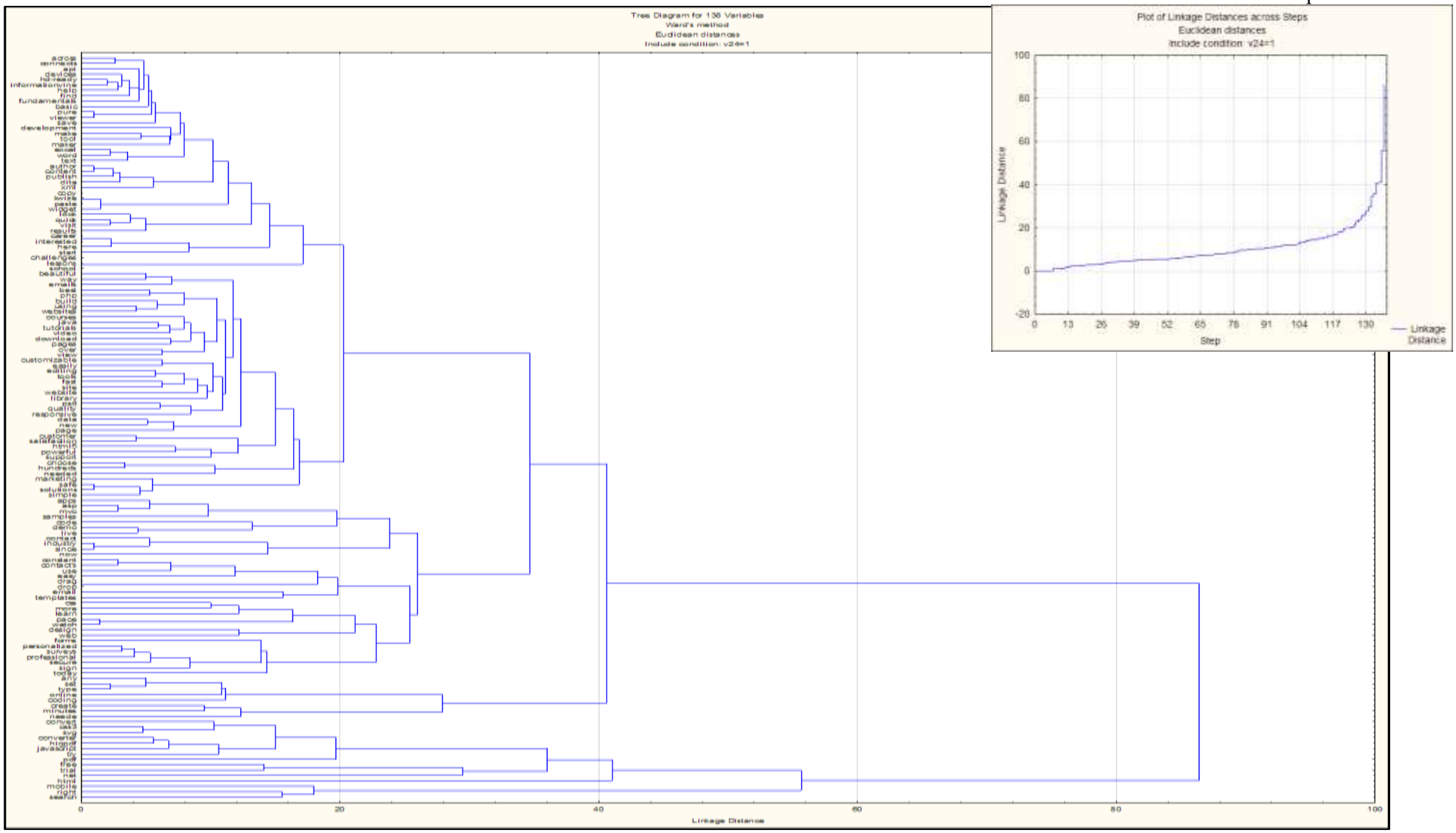

Рис. 3 – Вариант 2. Кластерный анализ поисковых запросов по переменной PPC. Страна US

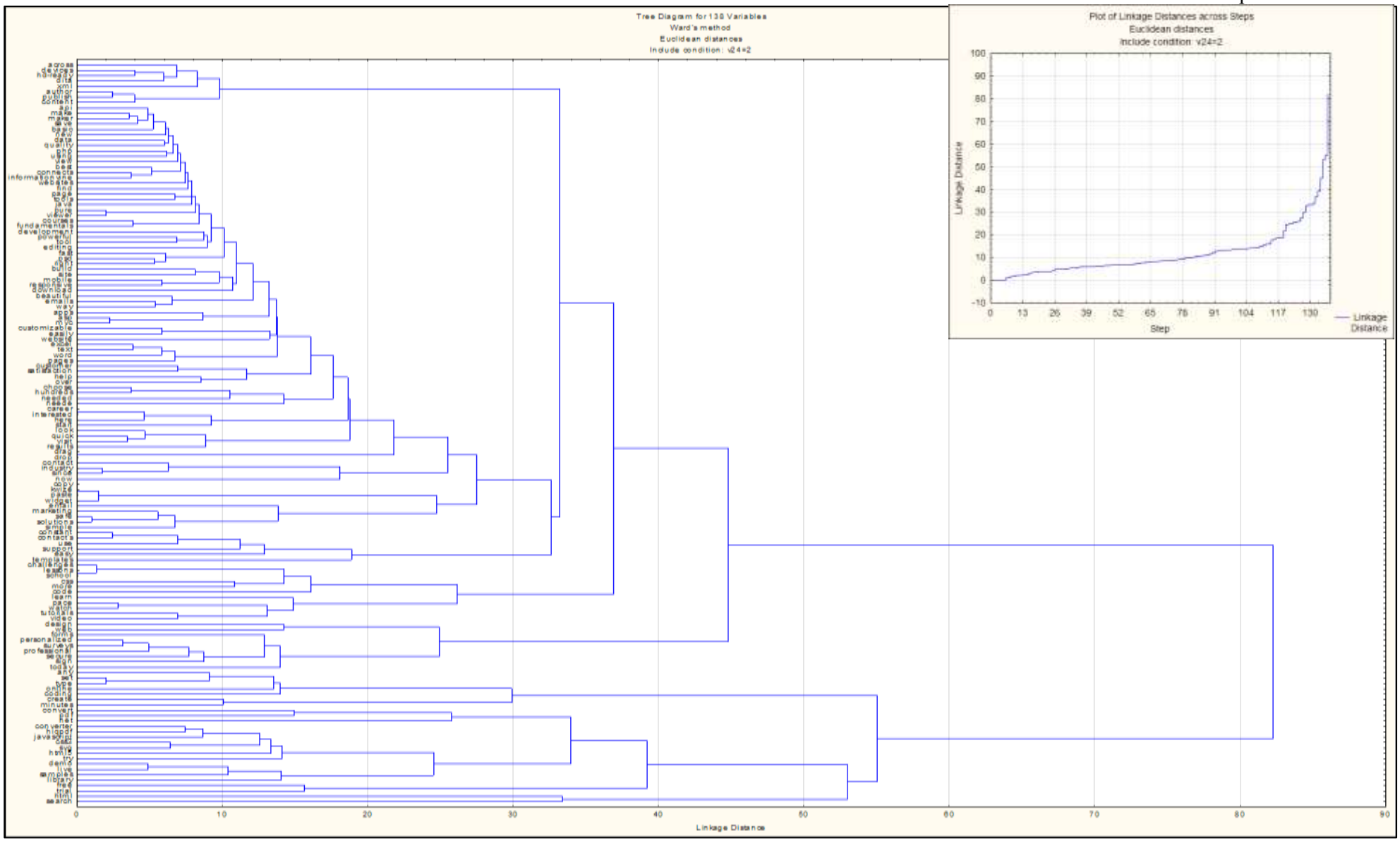

Рис. 4 – Вариант 2. Кластерный анализ поисковых запросов по переменной PPC. Страна US

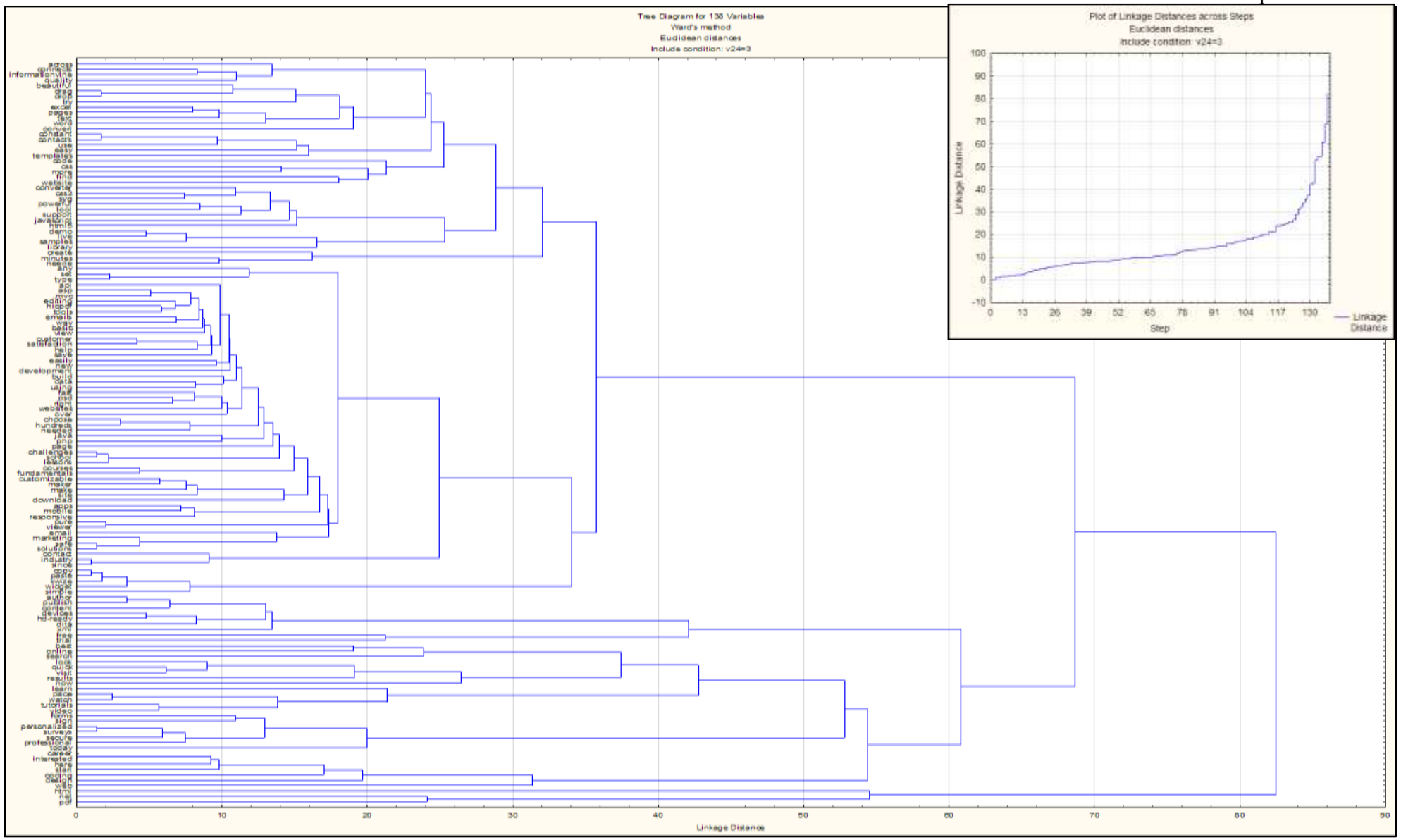

Рис. 5 – Вариант 2. Кластерный анализ поисковых запросов по переменной PPC. Страна US

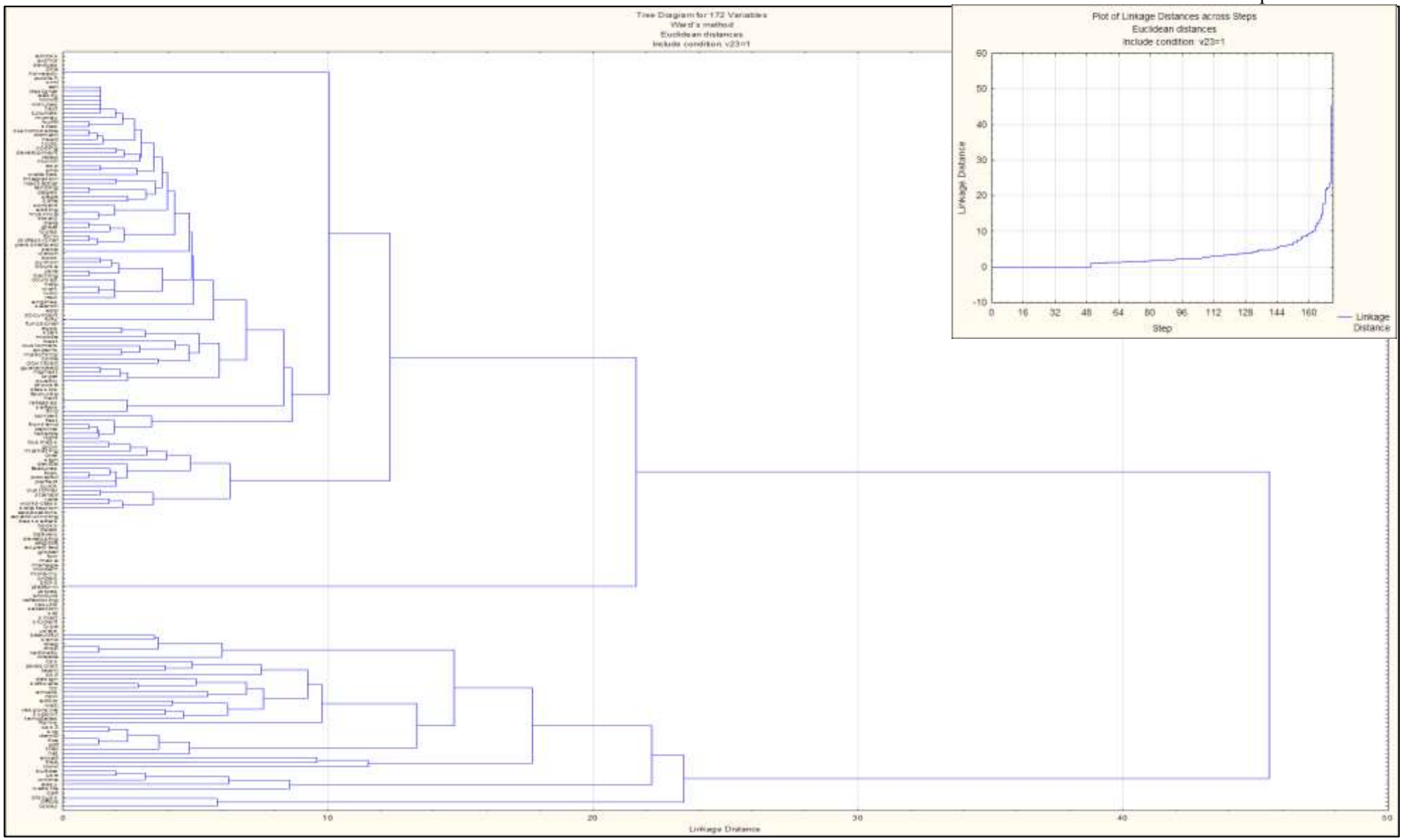

Рис. 6 – Вариант 3. Кластерный анализ поисковых запросов по переменной CPC. Страна UK

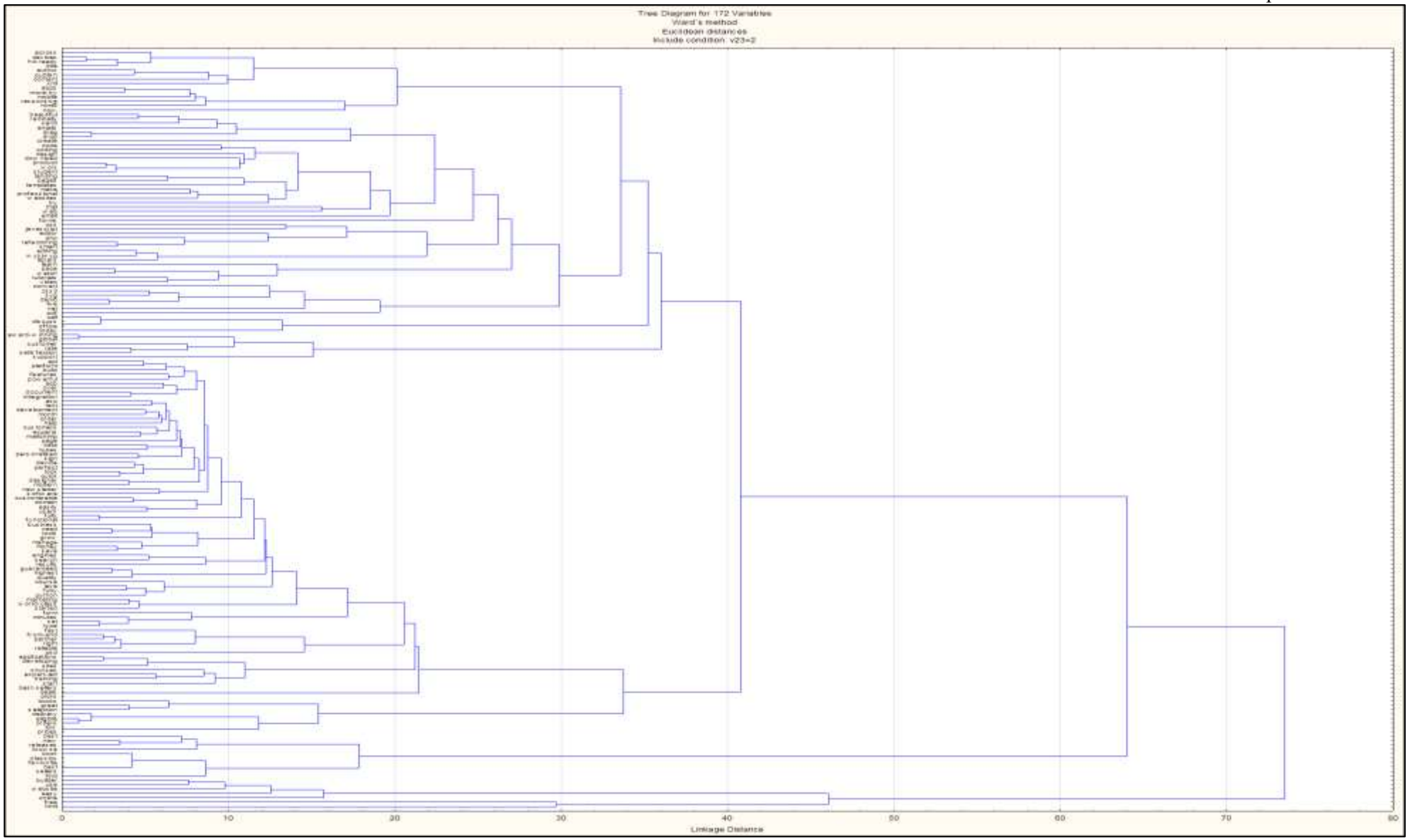

Рис. 7 – Вариант 3. Кластерный анализ поисковых запросов по переменной CPC. Страна UK

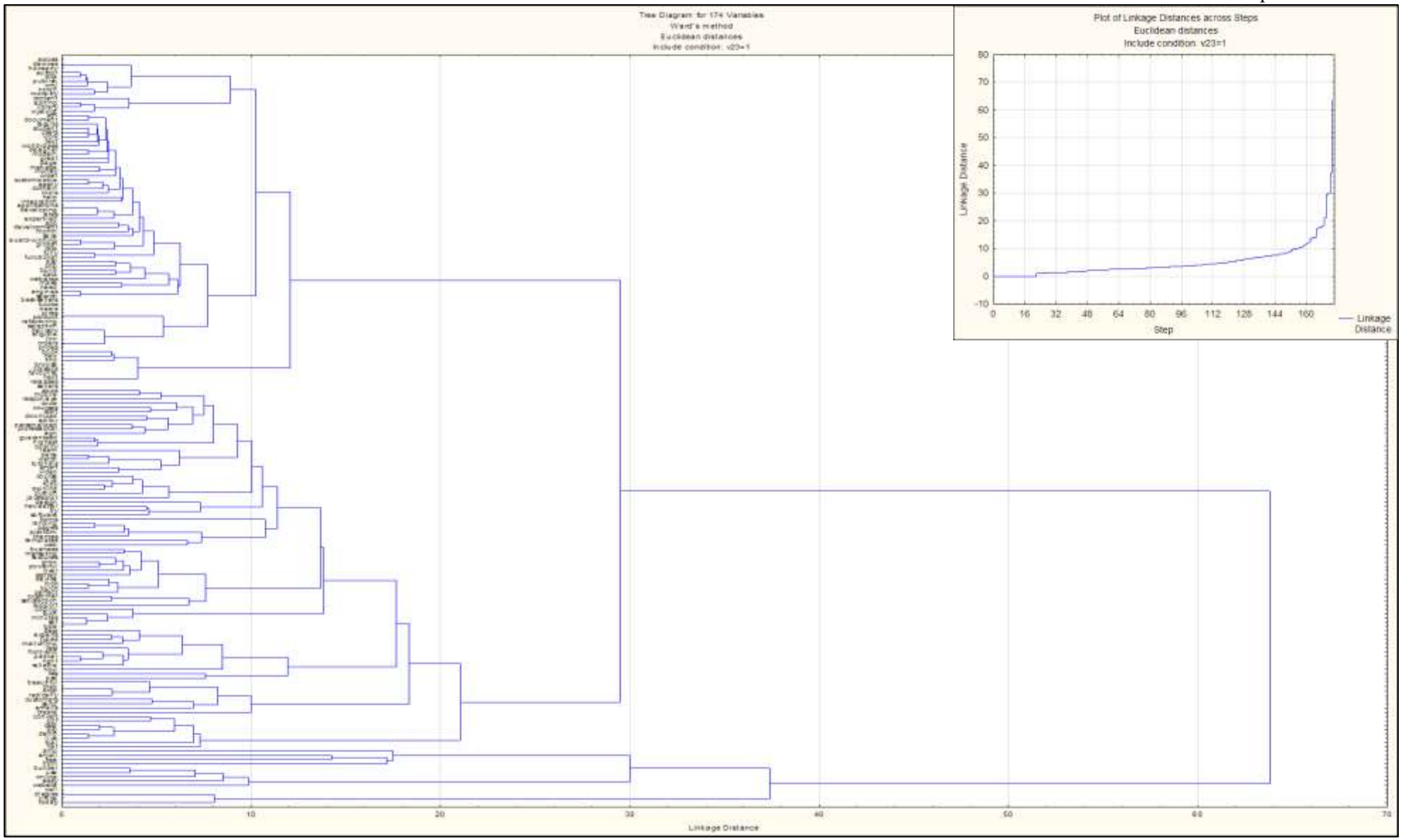

Рис. 8 – Вариант 4. Кластерный анализ поисковых запросов по переменной PPC. Страна UK

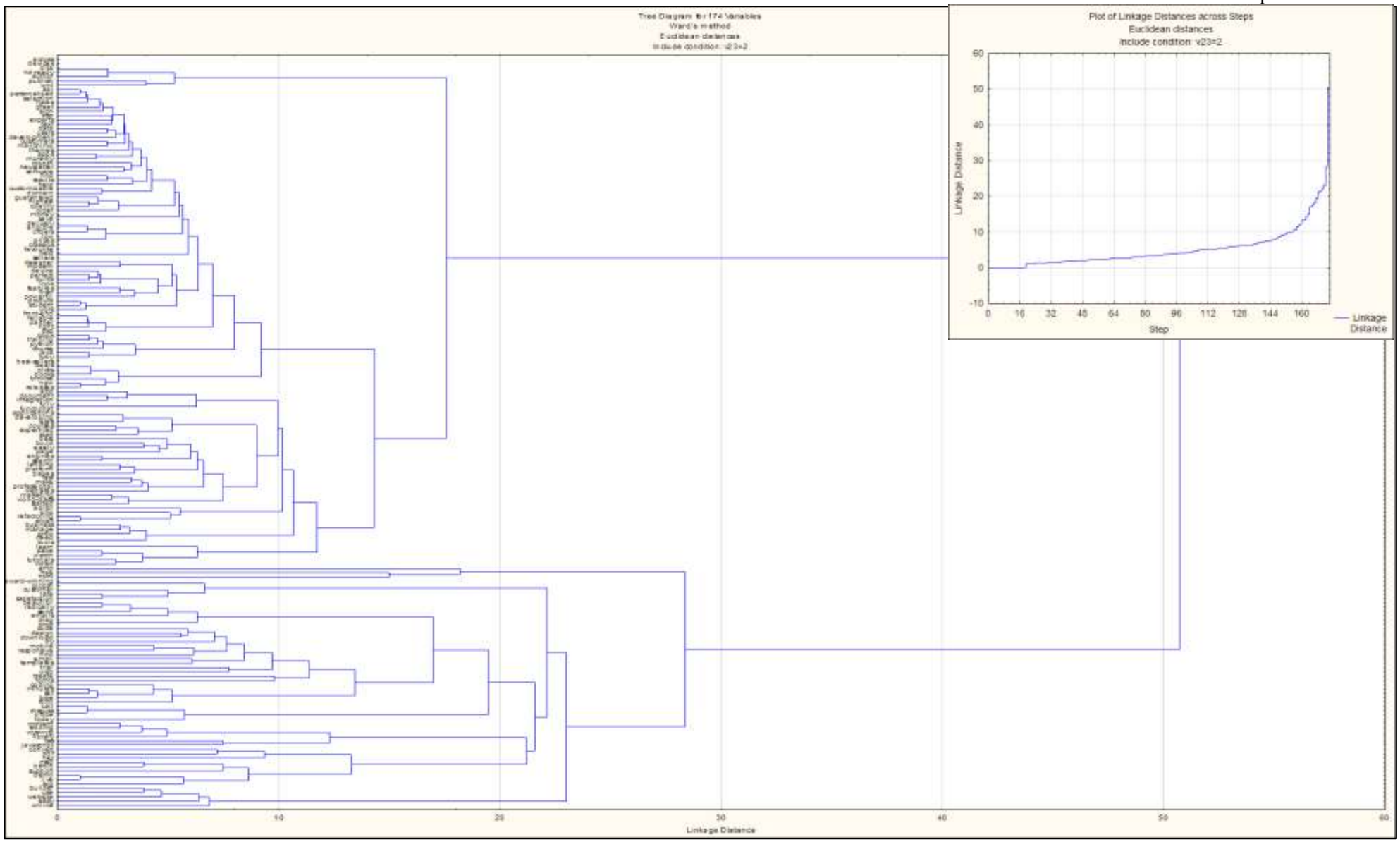

Рис. 9 – Вариант 4. Кластерный анализ поисковых запросов по переменной PPC. Страна UK

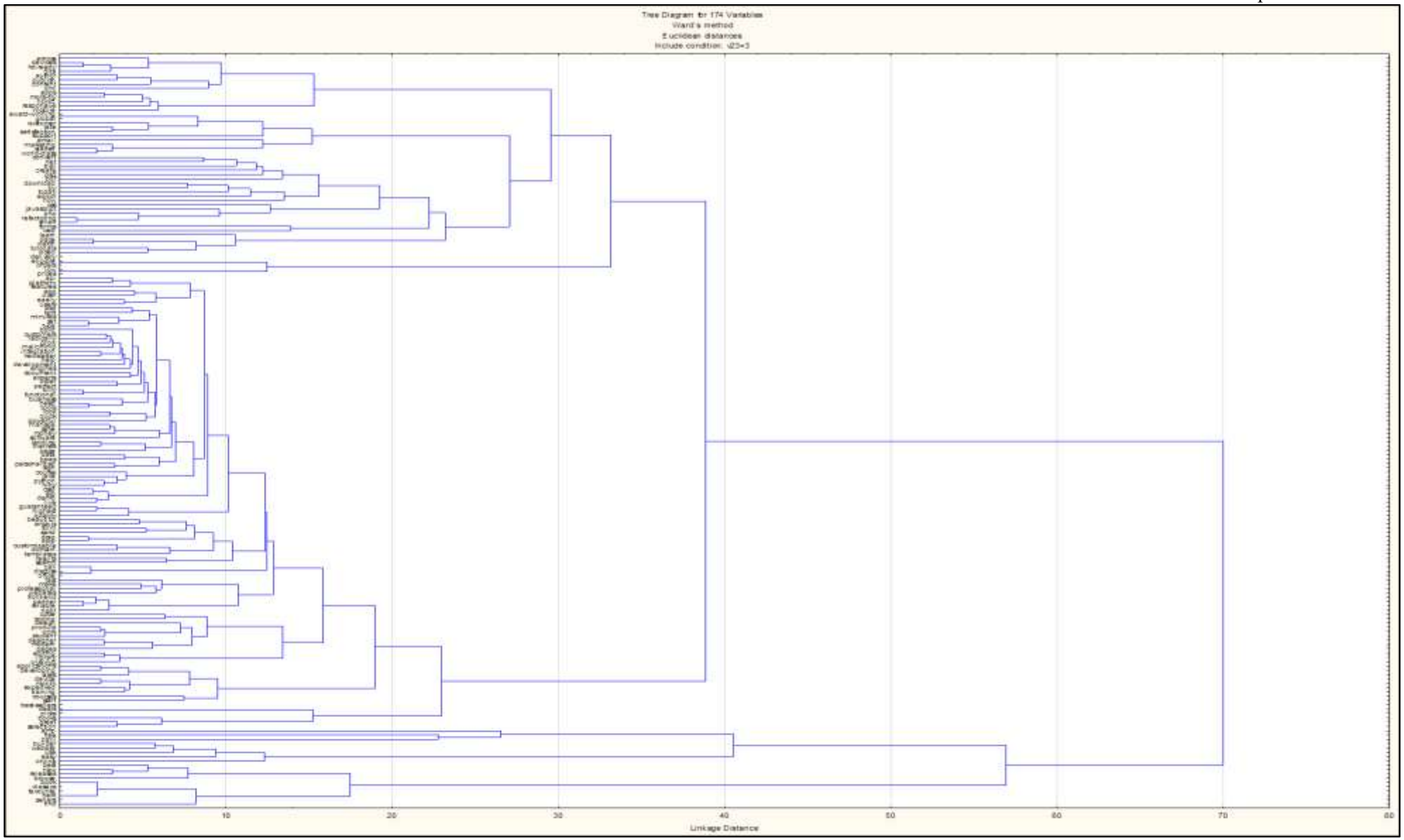

Рис. 10 – Вариант 4. Кластерный анализ поисковых запросов по переменной PPC. Страна US