**Иерархическая кластеризация** — метод кластерного анализа - разбиения заданной выборки объектов (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

**Цели**:

1. Понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»).

Число кластеров стараются сделать поменьше, чтобы легче было анализировать данные.

1. Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.

Важно обеспечить высокую степень сходства объектов внутри каждого кластера, а кластеров может быть сколько угодно.

1. Обнаружение новизны (англ. novelty detection). Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Среди алгоритмов иерархической кластеризации выделяются два основных типа: **восходящие** и **нисходящие** алгоритмы. Нисходящие алгоритмы работают по принципу «сверху-вниз»: в начале все объекты помещаются в один кластер, который затем разбивается на все более мелкие кластеры. Более распространены восходящие алгоритмы, которые в начале работы помещают каждый объект в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержаться в одном кластере. Таким образом строится система вложенных разбиений. Результаты таких алгоритмов обычно представляют в виде дерева – дендрограммы. Классический пример такого дерева – классификация животных и растений.

Способы **вычисления расстояния** между кластерами:

**Метод ближнего соседа** или **одиночная связь**. Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов. В результате работы этого метода кластеры представляются длинными "цепочками" или "волокнистыми" кластерами, "сцепленными вместе" только отдельными элементами, которые случайно оказались ближе остальных друг к другу. **Проблема** **метода** – высокая степень случайности (если точка, которая не должна была попасть в кластер оказалась слишком близко к текущей точке, то кластер будет построен неверно).

**Метод наиболее удаленных соседей** или **полная связь**. Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Метод хорошо использовать, когда объекты действительно происходят из различных "рощ". Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочечным", то этот метод не следует использовать. **Проблема** **метода** – чуствительность к шумам (если всего одна из точек одного кластера оказалась слишком далеко от другого кластера – считается что кластеры далеко друг от друга).

**Метод средней связи**— метод связывания; начинается аналогично методам одиночной или полной связи, однако критерием кластеризации является среднее расстояние от индивидуумов одного кластера до индивидуумов другого. Данный метод нивелирует недостатки предыдущих двух методов, таким образом, зачастую он является предпочтительным.

**Вычислительная сложность**. Сложность агломеративного иерархического алгоритма зависит от выбранной меры близости различия между кластерами и способа реализации идеи алгоритма. Известны реализации со следующими оценками вычислительной сложности:

1. Агломеративный (+) алгоритм
   1. Одиночная связь – .
   2. Полная связь – .
   3. Усреднение по группе – .
2. Дивизитивный(-) алгоритм.

При исполтзовании алгоритма *k*-средних соседей вычислительная сложность составляет .

К **недостатку** иерархических алгоритмов можно отнести систему полных разбиений, которая может являться излишней в контексте решаемой задачи.

Иерархическую кластеризацию можно применять для **TextMining**.

Наиболее простой и наглядный **пример** применения для кластеризации текстов. Пусть дано множество документов, необходимо тематически кластеризовать эти документы.

1. Выделить ключевые слова и tag’и.
2. Построить обратный индекс слов в документе (из слов предыдущего пункта).
3. Использовать каждое слово предыдущего документа как измерение. Таким образом каждый документ станет точкой многомерного пространства.
4. Применить один из алгоритмов иерархоческой клатеризации.

Во время выполнения кластеризации к tag’ам и к ключевым словм можно применять разные метрики (с различными коэффициентами) для более эффективной кластеризации.

Таким образом (теоретически), документы попавшие в один кластер будут принадледать одной теме.