Рекомендации по доработке магистерской диссертации “Algorithm for …”

Методологические предложения

Следует сузить тему исследования и формально поставить задачу построения ранжирующей модели как оптимизационную задачу. При этом предлагается использовать математическую нотацию вместо описательной. Следует указать в каких условиях предполагается использовать разрабатываемую модель.

Следует переработать терминологический словарь, сократив его. По возможности, убрать избыточные определения, например, «мобильный телефон» и ввести недостающие определения, например, «онтология», «ранжирующая модель» (вариант «алгоритм») с теми уточнениями, которые предполагает узкоспециальное исследование.

Следует показать обоснованность предлагаемой ранжирующей модели. Предлагается выполнить анализ свойств модели, анализ ошибки. Предлагается сравнить разработанную ранжирующую модель с аналогичной моделью, предложенной другими авторами согласно принятым в работе критериям.

Предлагается реструктурировать доклад. Примеры слайдов докладов, выполненных по рассматриваемой специальности студентов Сколтеха и МФТИ прилагаются.

Предлагается подготовить ответы на возможные вопросы экзаменационной комиссии. Примеры вопросов.

1. Каков вид целевой функции при оптимизации параметров модели?
2. Каковы критерии выбора признаков?
3. Является ли число параметров (либо число признаков) оптимальным при заданной целевой функции?
4. Почему (исходя из каких критериев качества) были выбраны именно эти признаки?
5. Каков минимально необходимый размер выборки для исследования свойств предлагаемой модели?
6. Исходя из каких критериев был выбран размер используемой выборки?
7. Какова обобщающая способность этой модели?
8. Является ли модель недообученной или переобученной?
9. Насколько устойчивы параметры модели к изменению состава выборки?
10. Проводилось ли сравнение с альтернативными моделями на данной выборке?

Содержательные предложения

Поскольку работа носит прикладной характер, следует уделить больше внимания анализу свойств предлагаемой ранжирующей модели (алгоритма). В частности, предлагается

1) переформулировать и уточнить цели вычислительного эксперимента,
2) явно описать выборку, по которой строилась модель,
3) описать предположения, допускаемые относительно выборки, ее статистические свойства,
4) описать модель в математической нотации, избегая использования исходного текста исполняемого кода,

5) описать способ разбиения выборки при выполнении процедуры скользящего контроля.

Свойства полученной модели предлагается проанализировать, в частности,

1) сравнить полученную модель с альтернативной, предложенной другими авторами (предпочтительно сравнение нескольких различных моделей), вычислив значения критериев качества на той же выборке с тем же способом скользящего контроля,
2) вычислить дисперсию функции ошибки на разбиении скользящего контроля,
3) вычислить дисперсии параметров модели, указать их значимость.

Предлагается проанализировать, не противоречит ли используемый способ оптимизации параметров модели (используемая целевая функция) принятым в работе критериям качества.

Предлагается оценить минимально необходимый объем выборки при заданной сложности модели. В рассматриваемом случае предлагается считать сложностью число параметров. Предлагается указать, является ли структура модели оптимальной (переобученной, недообученной) для данной выборки. Отсутствие анализа переобученности модели и оценки ее предсказательной способности считается значимым недостатком работы в данной области исследований.

Технические предложения

Важно. Не следует изымать из текста исходного кода, приведенного в Приложениях 8, 10 имена авторов (Pedro Matiello и соавт., copyright 2007-2009). Следует сохранить лицензионный текст и сослаться на источник этого кода. Также не рекомендуется включать исходный текст кода класса; см., например, Приложение 8 до строки 222, если этот класс до этого уже загружен командой строки 15. Не рекомендуется дважды включать один и тот же исходный текст (около 200 строк) в диссертационную работу. Источники:

1) класс digraph из А8 и А10 https://github.com/pmatiello/python-graph/blob/master/core/pygraph/classes/digraph.py
2) функция pagerank из А8 и А10 https://github.com/wting/python-graph/blob/master/core/pygraph/algorithms/pagerank.py

Важно. Рекомендуется проверить точность и достоверность высказываний диссертации. Рассмотрим следующее высказывание: “It is evident that large corpus of data and constant inflow of new information allows supervised algorithm to learn very effectively.” Во-первых, это неочевидно. Во-вторых, высказывание терминологически неопределенно. После уточнения терминов это высказывание может стать неверным. Пример. Рассмотрим модель логистической регрессии как частный случай исследуемой модели. Зафиксируем набор параметров, считая оценку параметров состоятельной, эффективной и несмещенной. Пусть исходная выборка является простой. Выборка пополняется, цитата: “constant inflow of new information”. Правдоподобие модели не увеличится в обоих случаях: как при сохранении, так и при изменении статистических свойств выборки вследствие ее пополнения.

Предлагается реорганизовать структуру работы. Сейчас она состоит из 14 основных разделов. Предлагается выделить три-четыре основные раздела и в каждом из них определить подразделы для получения иерархической структуры. Такая структура должна отделить важные сообщения от второстепенных.

Предлагается изъять из аннотации общие высказывания о развитии интернета и посвятить аннотацию описанию основных положений и результатов работы, как описано в руководстве по выполнению магистерской диссертации.

Предлагается переформулировать положения, называемые в аннотации и в тексте диссертации гипотезами, в виде целей и задач проекта. Предлагается избегать описательных постановок задачи.

Предлагается изъять из раздела “Выбор предмета исследования” общие фразы, например, “Information is one of the basic resources necessary for mankind” далее и сконцентрироваться на решаемой узкоспециальной задаче.

При описании работы поисковых машин следует указывать библиографические источники. Пример, где источник необходим: «For users’ convenience the largest search engines such as Google begun introducing verticalization algorithms for their services. Although, as the trend goes people prefer web-services specifically designed for vertical search rather than such extensions of the industry giants».

Предлагается изъять из текста работы высказывания, которые нельзя ни подтвердить, ни опровергнуть. Пример: "As the era of mobility flourished and the rhythm of life started to beat faster, people switched their preferences from horizontal search to vertical."

Предлагается из раздела “Определения” изъять все определения, которые не относятся к предмету исследований. Предлагается поставить ссылки на литературу, в которой приводятся определения основных терминов. Рекомендуется избегать включения в текст диссертации собственных определений, например, для мобильного телефона. Рекомендуется не включать тривиальные определения, и определения, не имеющие важного значения для исследования. Предлагается найти в тексте все высказывания, которые не имеют отношения к предмету исследований и изъять их.

Предлагается проверить корректность терминов, фактически используемых в работе, и изъять некорректные. Пример такого термина: "narrow data space”.

Научная новизна неочевидна в связи с тем, что предмет исследования не определен достаточно точно. Задача “построить систему вертикального поиска” является чрезмерно общей задачей, с учетом определения термина, приведенного в данной магистерской диссертации.

Раздел “Анализ существующих решений”. Следует избегать обобщающих высказываний, не подкрепленных доказательствами или библиографическими ссылками. Например, "all large search engines introduced verticalization algorithms apart from horizontal search”. Вариант: "The selection of verticals in this case is based on the fact that hundreds of thousands of users who queried Rihanna before us then proceeded to websites within one of these verticals. This algorithm is very efficient when the query is popular enough because the amount of information collected about users is sufficiently large.” Следует указать источник информации, привести результаты собственных вычислительных экспериментов, или сообщить, что приведенные факты являются оценкой автора.

Раздел “Гипотезы” рекомендуется переименовать в "Цели исследования”. Высказывания раздела сформулированы недостаточно определенно. Если сохранять подобную общность формулировок, то наиболее близкие исследования, опубликованы в

1) Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67,
2) Wolpert, D.H. (1996), "The Lack of A Priori Distinctions between Learning Algorithms", Neural Computation, pp. 1341-1390,

что уводит исследования из практической направленности в теоретическую.

Предлагается перенести исходные тексты программ в приложения. Предлагается для изложения использовать математическую нотацию или нотацию в формате подходящего языка UML.

Раздел 12.4 “Выбор параметров”. Рекомендуется объявить эту процедуру процедурой полного или частичного перебора и указать способ скользящего контроля, который использовался при отборе параметров.

Раздел 12.4 “Заключение”. Следует избегать неопределенностей в формулировках, например, "In some cases the algorithm extracts verticals better than human intelligence does." Рекомендуется привести численные оценки и указать, в каких случаях это высказывание верно.

В разделе “Экономические аспекты” рекомендуется заменить обзор глобального рынка на рекомендации по внедрению результатов работы и предполагаемую экономическую эффективность от внедрения.

Раздел “Заключение”: следует указать какие результаты были получены лично. Рекомендуется изъять общие соображение о важности темы исследования или перенести их в обзор литературы.

Следует оформить список литературы в соответствии с принятым стандартом. Следует упорядочить библиографические ссылки.

Следует проверить орфографию и пунктуацию всего текста диссертации.