**学号：2015200791**

**北京化工大学**

**硕士研究生学位论文开题报告**

**论文题目：基于**Fish-Search**搜索算法的乙烯能效知识发现方法研究及应用**

**学 院 名 称：信息科学与技术学院**

**专 业：计算机科学与技术**

**研究生姓名：高华超**

**导师姓名：耿志强**

**开题日期：2016年9月26日**

|  |  |  |
| --- | --- | --- |
| **考核**  **成绩** |  | |
| **审核**  **小组**  **成员**  **以及**  **职称** | **姓 名** | **职 称** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

目录

[1 课题来源及项目名称 1](#_Toc462834645)

[2 文献综述 1](#_Toc462834646)

[2.1 研究背景 1](#_Toc462834647)

[2.2 用到的主要技术的研究情况 2](#_Toc462834648)

[2.2.1 主题爬虫概述 2](#_Toc462834649)

[2.2.2多维概率分布概述 3](#_Toc462834650)

[2.3 主要方法 3](#_Toc462834651)

[2.3.1 主题爬虫方法 3](#_Toc462834652)

[2.3.2 多维概率分布求解方法 4](#_Toc462834653)

[2.3.3相似度度量方法 5](#_Toc462834654)

[2.3.4知识的表示方法 6](#_Toc462834655)

[2.4 课题创新 6](#_Toc462834656)

[2.5 参考文献 7](#_Toc462834657)

[3 研究计划 8](#_Toc462834658)

[3.1 选题的目的及意义 8](#_Toc462834659)

[3.2 主要研究内容 9](#_Toc462834660)

[3.3 研究方案 9](#_Toc462834661)

[3.3.1 技术方案 9](#_Toc462834662)

[3.3.2 实施方案所需要的条件 10](#_Toc462834663)

[3.3.3 拟解决的关键问题 10](#_Toc462834664)

[3.4 课题难点分析 11](#_Toc462834665)

[3.5 预期研究成果及创新点 11](#_Toc462834666)

[3.6 工作计划进度 12](#_Toc462834667)

# 1 课题来源及项目名称

课题来源：自选题目

项目名称：基于相似性搜索的能效评级方法研究及应用

# 2 文献综述

## 2.1 研究背景

乙烯是石油化工的主要产品之一,乙烯装置生产的三烯(乙烯、丙烯、丁二烯)和三苯(苯、甲苯、二甲苯)是石油化工工业的基础原料,乙烯作为石油化工的龙头装置,其能耗水平是衡量一个国家石油化工发展水平的重要标志。随着国家节能方针的逐步推进与深化,国内的石化企业正面临着越来越严峻的竞争和挑战。

乙烯行业的发展与国家工业的发展密切相关，其生产水平往往成为判断一个国家整体工业水平的重要标志之一［1］。据统计，2014年中国石化集团乙烯产量达10420 Kt•a-1，而乙烯燃动能耗（标油）为571.39 kg/ t[2]；总体来看，我国乙烯行业平均综合能耗比国外先进的同类装置偏高，因而国内乙烯行业仍存在较大的能效提升空间。综上，我们应该建立针对乙烯装置的能耗指标体系，并将其应用于乙烯工业，为降低石油化工的能源消耗树立积极的标杆[3]。能耗指标体系建立在对海量数据的分析之上，这包括：专家经验、操作规范、生产设计数据等。从这些海量信息中挖掘与乙烯能耗息息相关的重要因素，就单/多装置上升至整个乙烯工业，在多个层次框架上建立完整的乙烯能耗指标体系，对企业的乙烯能耗分布加以评估统计，从而寻求节能降耗的路径与指导方案，将量化的降耗目标分解落实。

在工业实际生产过程中普遍选取石油裂解生产乙烯[4]。生产过程中虽然采用了计算机控制系统如DCS实现实时监控，部分装置也实现了优化控制。但是上述控制仅仅只是针对过程参数的操作与控制，不能评价装置的能效。操作者根据自己的经验进行操作，无法知道装置的生产效率与能效是否达到最佳，如何调整相应的操作与管理参数才能提高与优化装置的能效等都是极具挑战性的问题。为此，有必要引入人工智能的方法，基于工业生产装置的操作数据，为工业乙烯生产提供用于指导装置操作运行的能效分析与预测的智能系统，提高装置的优化操作水平。

## 2.2 用到的主要技术的研究情况

### 2.2.1 主题爬虫概述

搜索引擎在国外的发展比国内要早一段时间，从上世纪90年代初期到现在才短短几十年的时间，搜索引擎迅速发展并成功的改变了人们的日常生活，为整个社会的进步以及互联网的发展产生了不可估量的影响作用。纵观搜索引擎历史，雅虎是国外第一个目录搜索引擎[5]，它是根据人工的方法来收集数据信息的；Larry Page和Sergey Brin创建的Google是国外最大的常用搜索引擎[6]，Google平均每天需要处理高达3亿次以上的搜索请求，为世界上包括中文在内的几十种常用语言提供着多种多样的服务和帮助。国内的搜索引擎起步相对来说比较晚，以李彦宏创建的百度为首的中文搜索引擎仍然为我国广大的用户提供了符合中文习惯的各式各样的搜索帮助，现如今百度已经成为国内使用最广泛的中文搜索引擎。

在传统的搜索引擎的基础上，一种新的主题搜索引擎逐渐得到了发展，那就是主题搜索引擎，主题搜索引擎可以面向各式各样的领域[7]。1994年出现了一个最原始的主题搜索引擎Fish Search System[8]，这个系统的问世在学术界引起了巨大的反响，使主题搜索引擎得到了很多人的关注。随后，主题搜索引擎的发展越来越迅速，其中一个比较有名的主题搜索引擎叫做CORA[9]，它可以自动的搜索并下载与计算机技术有关的论文，因为CORA具有优秀的可扩充性，因此可以将其思想应用到其他的领域中，CORA的出现，将主题搜索引擎的研究以及应用推向了另一个高峰；与CORA类似，Scirus也是一个可以为科研人员提供技术论文的主题搜索引擎[10]。在当今社会，主题搜索引擎的应用越来越多，尤其是在旅游搜索、房产搜索、酒店搜索、机票搜索等领域中使用越来越广泛。

主题搜索引擎的核心是主题爬虫，1993年MIT开发出了第一个领域概念意义上的爬虫程序WWWW（World Wide Web Wanderer），虽然后来的爬虫程序越来越复杂、功能越来越完善，但是源于这个爬虫的思想却一直延续到了现在[11]。在这个爬虫的基础上，逐渐涌现出了一大批非常优秀的主题爬虫：比如在IBM Focused Crawler项目中[12]，作者引出了一种全新的爬虫搜索策略，就是简易版的主题爬虫，它利用一组训练集来描述爬虫的相关主题，在这个主题爬虫中，作者创新性的提出了网页主题相关度，通过计算网页的相似度来确定网页与主题是否相关；无独有偶，印度理工大学也研究了一个Focus Project系统[13]，其思想也基本相似；国内比较出名的主题搜索引擎比如北大研制的“天网”搜索引擎[14]，这是国内第一个基于[网页](http://baike.baidu.com/view/828.htm)搜索的中英文搜索引擎，现如今又逐渐推出了个性化搜索和专业搜索业务，影响范围进一步扩大；严格来说，淘宝网也属于一个主题搜索引擎，它可以从其他网站平台抓取信息，为用户提供更加个性化的服务。

在主题爬虫研究领域，搜索策略的研究一直比较热门，总的来说，搜索策略主要有两种：基于内容的搜索策略和基于链接的搜索策略[15]。

主题爬虫的另一个研究热点是对主题相关度的研究，比如向量空间模型（VSM）、布尔模型、基于本体（Ontology）的文本相似度计算等。国内的研究主要有张焕炯等提出的使用汉明距离的方法进行相关性判断[16]；余刚等提出的基于词语语义的相关性判断方法[17]；潘谦红等提出的基于属性论的相关性判断方法[18]；张志刚等人提出的按照网页结构的不同进行内容抽取的方法[19]。

### 2.2.2多维概率分布概述

每个乙烯能效指标都具有多方面的影响因素，为全面了解其统计规律，需要从多个角度对它进行定义和描述。例如某一个指标可能被蒸汽、水、电所影响。在判别哪些因素对指标的影响最大时，可以看作是组合概率的问题，即某些因素对某一特定指标的影响。在数学书上，这些都可以采用联合概率分布进行求解。但是利用多维联合概率分布理论进行此类不确定性分析计算的研究较少，主要原因是多维联合分布的显示表达式不易获得，只有正态分布、对数正态分布等少数几种分布可以推导出多维联合分布的解析表达式，对其他分布的多维解析求解则相当困难。

目前，国内外多维联合分布的计算方法研究主要集中于边缘，分布为正态分布、对数正态分布、Gumbel分布、指数分布等, 但这些方法大多局限于二维分布, 对三维以上联合分布的求解较为困难[20]。

## 2.3 主要方法

### 2.3.1 主题爬虫方法

2.2.1节介绍了主题爬虫的大致的研究范围和情况，而关于主题爬虫的相关算法种类颇多，在这里介绍几类主题爬虫方法。在主题爬虫中使用最广泛的搜索策略是基于内容的搜索策略和基于链接的搜索策略。

1、基于内容的搜索策略

基于内容的搜索策略只考虑网页页面的正文与主题的关系，基本不考虑链接之间的内在关系。这种搜索策略的典型代表有Fish-Search和Shark-Search，这两种算法是在仿生学的基础上将爬虫比作鱼群在水中的觅食行为，将爬虫比作鱼群，将网页比作食物，相关网页之间的URL比作觅食的行为，如果有URL则说明觅食行为成立[21]。每次当鱼发现食物（相关网页）时，鱼就会进行繁殖，扩大鱼群的数量；当鱼群没有找到食物（相关网页）时，鱼就会饿死，减少鱼群的数量，直到鱼群逐渐消失。这两种算法的基本思想类似：

1. 从初始种子中选取一个起始点，并打开对应的页面；
2. 提取网页的内容，根据文字内容计算该网页与主题的相似度；
3. 如果该网页的相似度大于阈值，则提取网页中链接；若相似度小于阈值，舍弃该网页；
4. 将等待队列中的链接按照一定的规则进行排序，优先访问靠前的网页，直至访问结束。

其实Shark-Search是Fish-Search的改进版本，改进部分主要是相似性计算模块。Fish-Search计算出来的相似度只能是1、0.5、0其中的一个，这就限制了网页和主题之间的相关性判断，比如同样优先级的网页可能会因为等待队列中的网页过多而不能得到及时的爬取，而这个网页可能是一个非常优秀的相关网页，这就会使爬虫效率降低；Shark-Search计算出来的相似度取值范围是（0,1），参考了网页中的锚文本及其上下文内容，并且引入了向量空间模型方法，使得相似度的计算比Fish-Search算法更加精确。

2、基于链接的搜索策略

在这种搜索策略中的代表算法有PageRank和HITS算法[22]。

PageRank算法基本思路是迭代的思想，在计算网页的PageRank之前，每个网页都设定一个初始值1.0，由于互联网上各个网页之间会存在着环的联系，所以计算网页的PageRank值是一个不断迭代的过程，随着迭代次数的不断增加，PageRank值会逐渐趋向于一个稳定的值。按照Larry Page的实验，对于3亿左右的链接，大约进行52次迭代就可以求得一个网页趋于稳定的PageRank值；对于1.6亿左右的链接，大约进行45次迭代就可以求得一个网页趋于稳定的PageRank值[22]，由此可见， PageRank算法适用于大规模的网页之间的相关性计算。

与PageRank算法齐名的还有一个HITS算法，考虑到用户访问网页不都是随机的，而是具有一定需求的，在此基础上，HITS算法提出了两个概念：Authority页面和Hub页面。Authority页面指的是那些已经被认证的高质量的网站页面、各大门户网站首页、权威网站首页等；Hub页面指的是拥有较高出入度的网页。每一个网页都设定一个Authority值和一个Hub值，如果网页的入度相对较高，则这个网页的Authority值就会相对越高；一个网页的出度相对较高，则这个网页的Hub值就会相对较高。类似于PageRank算法，HITS算法中的Authority值和Hub值也都有一个初始值，每次迭代后值都会进行归一化处理，经过多次迭代后Authority值和Hub值都会收敛于一个趋向稳定的值，而这个值就是这个网页最终的Authority值和Hub值。

由于基于链接的搜索策略并没有考虑网页页面上的内容，所以在主题爬虫的搜索过程中，容易发生“主题偏移”的现象。

### 2.3.2 多维概率分布求解方法

设是概率空间，是定义在上的取值于的向量值函数，如果对任意的，，则称为*n*维随机变量，称

，

为的联合分布函数。

设为*n* 维连续型随机变量, 即存在定义于上的非负数实值函数 （称为的概率密度），对于任意 , 随机变量的联合分布函数可表示为

记落在微分区间(*,* ] *=*(*,* +*;… ; ,* +)的概率为

+++…+

通过比较，即得d （2-1）

通过公式（1）可以得到，落在微分区间(*,* ] *=*(*,* +*;… ; ,* +)的概率为[23]。

综上，求解多维联合分布的关键是求出其概率密度函数。在这里介绍一种应用比较广泛的多维联合分布计算方法：将原始数据变换为正态的Moran方法。

由于原始变量的多维联合分布的解析求解比较困难，通常将原始变量转换为正态分布变量，继而求解多维正态分布概率。多维正态分布的密度函数如下：

（2-2）

式中：为正态分布变量；Σ为协方差矩阵，，。

采用Moran方法计算多维联合分布的一个关键问题是采用何种正态化变换工具。常用的正态化变换工具主要有Box-Cox 变换和多项式正态变换(PNT)。

对一组原始样本，Box-Cox变换可通过式（2-3）将其变换成近似正态分布样本

（2-3）

式中:是原始样本； 是变换后的样本；是变换参数；服从正态分布*N(μ, σ)*；*μ*为均值；*σ*为标准差。

### 2.3.3相似度度量方法

相似度度量（Similarity），即计算个体间的相似程度，相似度度量的值越小，说明个体间相似度越小，相似度的值越大说明个体差异越大[24~27]。常用的判断向量间差异的方法主要有欧氏距离和余弦相似度两种方法。欧式距离是常用的距离度量，两个向量间距离越大，说明差异越大。余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小[28]。余弦值越接近1，就表明夹角越接近0度，也就是两个向量越相似，这就叫"余弦相似性"[29]。

两个向量a,b的夹角很大可以说a向量和b向量有很低的的相似性，或者说a和b向量代表的文本基本不相似。那么是否可以用两个向量的夹角大小的函数值来计算个体的相似度呢？向量空间余弦相似度理论就是基于上述来计算个体相似度的一种方法[30~32]。下面做详细的推理过程分析。

三角形中边a和b的夹角的余弦计算公式为：

 （2-4）

在向量表示的三角形中，假设a向量是（x1, y1），b向量是(x2, y2)，那么可以将余弦定理改写成下面的形式：向量a和向量b的夹角的余弦计算如下：

 （2-5）

如果向量a和b不是二维而是n维，上述余弦的计算法仍然正确。假定a和b是两个n维向量，则a与b的夹角的余弦等于：

 （2-6）

### 2.3.4知识的表示方法

知识表示是人工智能的中心研究课题之一，它不仅是知识工程的关键问题，而且还是其他AI 研究中的重要问题。它的首要任务是能够准确的表达人类知识，但是目前在把传统的知识表示方法应用于模糊推理的专家系统中时，又面临着不能准确表达领域知识等问题。因此，很多文献提出了框架结构和模糊产生式相结合的方法。这种方法，把前提中的各个框架的作用视为同等重要，但在不少领域，前提中每个框架的作用是不一样的，这种方法就不能满足实际的需要。另外，每一条知识的可信程度是不一样的，因此，我们又加入了可信度。此外，这种方法仅判断前提的各个框架存在与否，而不判断框架的发生强度。例如，在矿产判断中，出现某种矿的矿脉的数目只有少数几个与数目很多，对于矿藏的判断是不一样的，即二者发生的强度不一样，因此，我们又提出了强度系数的概念，给出了各个框架发生的强度。在模糊推理中，一般的推理算法，都是推到终结点，其可信度超过规定的阈值，则成立，否则为不成立。

下面给出重要度和强度系数的定义。

（1）重要度

定义 1 一个规则的前提条件中and的个数n 称为分支数。

定义 2 每个分支赋以一个实数 IM（0 < *IM* ≤1），称为该分支的重要度，它满足

其中，n为该规则的分支数。

它们的意义是： 表示该规则在n个前提中的第i个分支的重要程度。例如，如果

≥ ，表示该规则的结论要成立，则第 i个分支比第 j个分支重要。它是在建立专家系统时，由领域专家给出。

（2）定义3 对于每个断言，由于其存在的时间长短或连续与否以及发生量的多少等强度因

素，会对断言的隶属度有影响，因此赋以一个实数 FT，称为强度系数，有。它的意义是，FT＝1 时，强度系数 FT 对规则的隶属度没有影响，当*FT* ≥1时，强度系数FT 将对规则的隶属度其作用，即规则的隶属度将因FT 的变化而变化。

## 2.4 课题创新

乙烯能耗水平是衡量一个国家石油化工发展水平的重要标志，我们要引入人工智能的方法，基于工业生产装置的操作数据，为工业乙烯生产提供用于指导装置操作运行的能效分析与预测的智能系统，提高装置的优化操作水平。

课题第一部分工作是采用改进的Fish-Search算法，利用合适的相似度算法，从数据库中获取所需数据；第二部分工作则是基于第一点，挖掘到的指标数据都可以看作知识，利用多维概率分布对大量能效数据进行规则的挖掘，然后进行等级划分，对所输入的数据找到相应等级。第三点是基于第一部分获得的同一装置的数据，利用合适的数据融合算法对输入的监测数据进行产量及能耗的预测。

## 2.5 参考文献

[1] Zong Yangong（宗言恭）. The status of China’s ethylene industry and its development[J].China Petroleum and Chemical Economic Analysis（中国石油和化工经济分析），2006,13:47-54

[2] 马国锋，徐跃华，郭新. 中国石化2014年乙烯业务述评[J].乙烯工业.2015, 27(1):1-5.

[3] 李涛. 乙烯生产原料的发展状况分析[J]. 石油化工技术与经济, 2005, 21(5):12-17

[4] 杨旭弘. 浅论乙烯的市场与生产[J]. 乙烯工业, 2000, (2):52-56.

[5] [H Dong](http://xueshu.baidu.com/s?wd=author%3A%28Hai%20Dong%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson), [FK Hussain](http://xueshu.baidu.com/s?wd=author%3A%28Hussain%2C%20F.K.%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson). [Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery](http://xueshu.baidu.com/s?wd=paperuri%3A%28de0792e32d4fe7dfcc66a276bdb0b419%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Ficp.jsp%3Farnumber%3D6384736&ie=utf-8) [J]. [IEEE Transactions on Industrial Informatics](http://xueshu.baidu.com/s?wd=journaluri%3A%28eb06b52fb8280f62%29%20%E3%80%8AIEEE%20Transactions%20on%20Industrial%20Informatics%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2014, 10(2): 1616-1626

[6] [王飞红](http://xueshu.baidu.com/s?wd=authoruri%3A%28b80d7ef8af6f27f8%29%20author%3A%28%E7%8E%8B%E9%A3%9E%E7%BA%A2%29%20%E8%B4%B5%E5%B7%9E%E5%B8%88%E8%8C%83%E5%AD%A6%E9%99%A2%E6%95%B0%E5%AD%A6%E4%B8%8E%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E5%AD%A6%E9%99%A2&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited), [丁泽发](http://xueshu.baidu.com/s?wd=authoruri%3A%28b91fca2eae460926%29%20author%3A%28%E4%B8%81%E6%B3%BD%E5%8F%91%29%20%E8%B4%B5%E5%B7%9E%E5%B8%88%E8%8C%83%E5%AD%A6%E9%99%A2%E6%95%B0%E5%AD%A6%E4%B8%8E%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E5%AD%A6%E9%99%A2&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited). [基于Lucene的垂直搜索引擎设计与实现](http://xueshu.baidu.com/s?wd=paperuri%3A%284c54f1c3072f1159630d6b778a17067a%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.cqvip.com%2FQK%2F80675A%2F201405%2F49288623.html&ie=utf-8)[J]. [电子技术与软件工程](http://xueshu.baidu.com/s?wd=journaluri%3A%28d66c4aaf8fbe8b2f%29%20%E3%80%8A%E7%94%B5%E5%AD%90%E6%8A%80%E6%9C%AF%E4%B8%8E%E8%BD%AF%E4%BB%B6%E5%B7%A5%E7%A8%8B%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2014(5): 206-208

[7] Kevin Wallsten. An Analysis of the Relationship between Mainstream Media and Political Blogs [J]. Agenda Setting and Blogosphere, 2007, 24(6): 567-667

[8] [YJ Du](http://xueshu.baidu.com/s?wd=author%3A%28YaJun%20Du%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson), [QQ Pen](http://xueshu.baidu.com/s?wd=author%3A%28QiangQiang%20Pen%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson), [ZQ Gao](http://xueshu.baidu.com/s?wd=author%3A%28ZhaoQiong%20Gao%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson). [A topic-specific crawling strategy based on semantics similarity](http://xueshu.baidu.com/s?wd=paperuri%3A%28d54f1a36fbf83c3b56980f9afd613323%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.sciencedirect.com%2Fscience%2Farticle%2Fpii%2FS0169023X13001079&ie=utf-8) [J]. [Data & Knowledge Engineering](http://xueshu.baidu.com/s?wd=journaluri%3A%281867581007d23e1d%29%20%E3%80%8AData%20%26%20Knowledge%20Engineering%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2013, 88(6): 75-93

[9] 池勇敏，郝泳涛. 分布式主题爬虫的设计与实现[J]. 计算机应用与软件, 2010, 27(12): 135-138

[10] [韩琳](http://xueshu.baidu.com/s?wd=author%3A%28%E9%9F%A9%E7%90%B3%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson), [基于贝叶斯主题爬虫的研究与实现](http://xueshu.baidu.com/s?wd=paperuri%3A%2865ec5f1d9bf7dfc7f15d6d7e4b3b39cd%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fcdmd.cnki.com.cn%2FArticle%2FCDMD-10005-1016701305.htm&ie=utf-8)[D]. 北京: 北京工业大学, 2015

[12] [朱学芳](http://xueshu.baidu.com/usercenter/data/author?cmd=authoruri&wd=authoruri%3A%2873ba22ebb04956b0%29%20author%3A%28%E6%9C%B1%E5%AD%A6%E8%8A%B3%29%20%E5%8D%97%E4%BA%AC%E5%A4%A7%E5%AD%A6%E4%BF%A1%E6%81%AF%E7%AE%A1%E7%90%86%E7%B3%BB)，[韩占校](http://xueshu.baidu.com/usercenter/data/author?cmd=authoruri&wd=authoruri%3A%286c15b3f3afab0e0f%29%20author%3A%28%E9%9F%A9%E5%8D%A0%E6%A0%A1%29%20%E5%8D%97%E4%BA%AC%E5%A4%A7%E5%AD%A6%E4%BF%A1%E6%81%AF%E7%AE%A1%E7%90%86%E7%B3%BB%E3%80%81%E5%A4%9A%E5%AA%92%E4%BD%93%E4%BF%A1%E6%81%AF%E5%A4%84%E7%90%86%E7%A0%94%E7%A9%B6%E6%89%80). [基于P2P的分布式主题爬虫系统的设计与实现](http://xueshu.baidu.com/s?wd=paperuri%3A%289503858b35ea9006263b1030e5a5c02b%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.cqvip.com%2FQK%2F95888X%2F201003%2F34317338.html&ie=utf-8)[J]. [情报学报](http://xueshu.baidu.com/s?wd=journaluri%3A%2869240599bf9285aa%29%20%E3%80%8A%E6%83%85%E6%8A%A5%E5%AD%A6%E6%8A%A5%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2010, 29(3): 402-407

[13] [P Bedi](http://xueshu.baidu.com/s?wd=author%3A%28Punam%20Bedi%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson), [A Thukral](http://xueshu.baidu.com/s?wd=author%3A%28Anjali%20Thukral%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson), [H Banati](http://xueshu.baidu.com/s?wd=author%3A%28Hema%20Banati%29%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson). [Focused crawling of tagged web resources using ontology](http://xueshu.baidu.com/s?wd=paperuri%3A%28a0cbb2a52bba166943964fa1cd64f55f%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.sciencedirect.com%2Fscience%2Farticle%2Fpii%2FS0045790612001759&ie=utf-8)[J]. [Computers & Electrical Engineering](http://xueshu.baidu.com/s?wd=journaluri%3A%28e2ccf5f2397f0261%29%20%E3%80%8AComputers%20%26%20Electrical%20Engineering%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2013, 39(2): 613-628

[14] [茹立云](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e8%8c%b9%e7%ab%8b%e4%ba%91%22), [李智超](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e6%9d%8e%e6%99%ba%e8%b6%85%22), [马少平](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e9%a9%ac%e5%b0%91%e5%b9%b3%22). 搜索引擎索引网页集合选取方法研究[J]. 计算机研究与发展, 2014, 51(10): 2239-2247

[15] [朱庆生](http://xueshu.baidu.com/usercenter/data/author?cmd=authoruri&wd=authoruri%3A%286db6cb3ed67b4ad4%29%20author%3A%28%E6%9C%B1%E5%BA%86%E7%94%9F%29%20%E9%87%8D%E5%BA%86%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2), [徐宁](http://xueshu.baidu.com/s?wd=authoruri%3A%2848fc1f75a4ca824d%29%20author%3A%28%E5%BE%90%E5%AE%81%29%20%E9%87%8D%E5%BA%86%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2%E8%BD%AF%E4%BB%B6%E7%90%86%E8%AE%BA%E4%B8%8E%E6%8A%80%E6%9C%AF%E9%87%8D%E5%BA%86%E5%B8%82%E9%87%8D%E7%82%B9%E5%AE%9E%E9%AA%8C%E5%AE%A4&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited), [周瑜](http://xueshu.baidu.com/s?wd=authoruri%3A%282be470a9a4468988%29%20author%3A%28%E5%91%A8%E7%91%9C%29%20%E9%87%8D%E5%BA%86%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2%E8%BD%AF%E4%BB%B6%E7%90%86%E8%AE%BA%E4%B8%8E%E6%8A%80%E6%9C%AF%E9%87%8D%E5%BA%86%E5%B8%82%E9%87%8D%E7%82%B9%E5%AE%9E%E9%AA%8C%E5%AE%A4&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited). [一种基于链接和内容分析的自适应主题爬虫算法](http://xueshu.baidu.com/s?wd=paperuri%3A%289cc28e1b7826c7aa241f0f3a04d711b0%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.cqvip.com%2FQK%2F97264X%2F201509%2F666059554.html&ie=utf-8)[J]. [计算机与现代化](http://xueshu.baidu.com/s?wd=journaluri%3A%284032d2f3794fc22%29%20%E3%80%8A%E8%AE%A1%E7%AE%97%E6%9C%BA%E4%B8%8E%E7%8E%B0%E4%BB%A3%E5%8C%96%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2015(9): 77-80

[16] [张焕炯](http://xueshu.baidu.com/usercenter/data/author?cmd=authoruri&wd=authoruri%3A%28ee928012bfc30a19%29%20author%3A%28%E5%BC%A0%E7%84%95%E7%82%AF%29%20%E6%9D%AD%E5%B7%9E%E7%94%B5%E5%AD%90%E7%A7%91%E6%8A%80%E5%A4%A7%E5%AD%A6). [基于汉明距离的文本相似度计算](http://xueshu.baidu.com/s?wd=paperuri%3A%281be782ab939a9c2b74b70ac0d1ee690f%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.cqvip.com%2FQK%2F91690X%2F200119%2F5474052.html&ie=utf-8)[J]. [计算机工程与应用](http://xueshu.baidu.com/s?wd=journaluri%3A%2876208619298e4f85%29%20%E3%80%8A%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%B7%A5%E7%A8%8B%E4%B8%8E%E5%BA%94%E7%94%A8%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2001, 37(19): 21-22

[17] [余刚](http://xueshu.baidu.com/s?wd=authoruri%3A%2820058f37d5290f4d%29%20author%3A%28%E4%BD%99%E5%88%9A%29%20%E9%87%8D%E5%BA%86%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited), [裴仰军](http://xueshu.baidu.com/usercenter/data/author?cmd=authoruri&wd=authoruri%3A%28fe03d2855ddab3fe%29%20author%3A%28%E8%A3%B4%E4%BB%B0%E5%86%9B%29%20%E9%87%8D%E5%BA%86%E7%A7%91%E6%8A%80%E5%AD%A6%E9%99%A2), [朱征宇](http://xueshu.baidu.com/usercenter/data/author?cmd=authoruri&wd=authoruri%3A%287e9e98c7d771361f%29%20author%3A%28%E6%9C%B1%E5%BE%81%E5%AE%87%29%20%E9%87%8D%E5%BA%86%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2)，陈华月. [基于词汇语义计算的文本相似度研究](http://xueshu.baidu.com/s?wd=paperuri%3A%28c66e4a4ec615b6c04554daa0b39cc62b%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.cqvip.com%2FMain%2FDetail.aspx%3Fid%3D21163104&ie=utf-8)[J]. [计算机工程与设计](http://xueshu.baidu.com/s?wd=journaluri%3A%288fbde9933399b238%29%20%E3%80%8A%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%B7%A5%E7%A8%8B%E4%B8%8E%E8%AE%BE%E8%AE%A1%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2006, 27(2): 241-244

[18] [潘谦红](http://xueshu.baidu.com/usercenter/data/author?cmd=authoruri&wd=authoruri%3A%2847feef5db9b94c4a%29%20author%3A%28%E6%BD%98%E8%B0%A6%E7%BA%A2%29%20%E4%B8%AD%E5%9B%BD%E7%A7%91%E5%AD%A6%E9%99%A2%E8%AE%A1%E7%AE%97%E6%8A%80%E6%9C%AF%E7%A0%94%E7%A9%B6%E6%89%80), [王炬](http://xueshu.baidu.com/s?wd=authoruri%3A%28475b5600333464fc%29%20author%3A%28%E7%8E%8B%E7%82%AC%29%20%E4%B8%AD%E5%9B%BD%E7%A7%91%E5%AD%A6%E9%99%A2%E8%AE%A1%E7%AE%97%E6%8A%80%E6%9C%AF%E7%A0%94%E7%A9%B6%E6%89%80&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited). [基于属性论的文本相似度计算](http://xueshu.baidu.com/s?wd=paperuri%3A%28f4dda3d277cf889c4542e63de093d677%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fwww.cqvip.com%2FMain%2FDetail.aspx%3Fid%3D3512757&ie=utf-8)[J]. [计算机学报](http://xueshu.baidu.com/s?wd=journaluri%3A%28bfb0f623432b99f6%29%20%E3%80%8A%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E6%8A%A5%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 1999, 22(6): 651-655

[19] [张志刚](http://xueshu.baidu.com/usercenter/data/author?cmd=authoruri&wd=authoruri%3A%2877093ab1234659bc%29%20author%3A%28%E5%BC%A0%E5%BF%97%E5%88%9A%29%20%E5%8C%97%E4%BA%AC%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E6%8A%80%E6%9C%AF%E7%B3%BB), [陈静](http://xueshu.baidu.com/s?wd=authoruri%3A%28a9858b44f89e6c9%29%20author%3A%28%E9%99%88%E9%9D%99%29%20%E5%8C%97%E4%BA%AC%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A7%91%E5%AD%A6%E6%8A%80%E6%9C%AF%E7%B3%BB%E7%BD%91%E7%BB%9C%E4%B8%8E%E5%88%86%E5%B8%83%E5%BC%8F%E7%B3%BB%E7%BB%9F%E5%AE%9E%E9%AA%8C%E5%AE%A4&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited), [李晓明](http://xueshu.baidu.com/s?wd=authoruri%3A%286be8c124d5a780aa%29%20author%3A%28%E6%9D%8E%E6%99%93%E6%98%8E%29%20%E5%8C%97%E4%BA%AC%E5%A4%A7%E5%AD%A6%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited). [一种HTML网页净化方法](http://xueshu.baidu.com/s?wd=paperuri%3A%28b25fe45779ea3ad94dc570246a6f0b9e%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fd.wanfangdata.com.cn%2FPeriodical%2Fqbxb200404001&ie=utf-8)[J]. [情报学报](http://xueshu.baidu.com/s?wd=journaluri%3A%2869240599bf9285aa%29%20%E3%80%8A%E6%83%85%E6%8A%A5%E5%AD%A6%E6%8A%A5%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 2004, 23(4): 387-393.

[20] 戴昌军, 梁忠民. 多维联合分布计算方法及其在水文中的应用[J]. 水利學報, 2006, 37(2): 160-165.

[21] 林海霞, 袁福永, 陈金森，刘俊峰. 一种改进的主题网络蜘蛛搜索算法[J]. 计算机工程与应用, 2007, 43(10): 174-176

[22] 张翔, 周明全, 李智杰，董丽丽. 基于PageRank与Bagging的主题爬虫研究[J]. 计算机工程与设计, 2010, 31(14): 3309-3312

[23] 陈朝舜. 微分法在求多维连续型随机变量函数的概率密度中的应用[J]. 渝州大学学报: 自然科学版, 2002, 19(1): 9-11.

[24] Ye J, “Cosine similarity measures for intuitionistic fuzzy sets and their applications,” Mathematical & Computer Modelling, 53(1-2):91-97,2011.

[25] Fawaz S. Al-Anzi, Dia AbuZeina, “Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing,” Journal of King Saud University - Computer and Information Sciences.

[26] Xia P, Zhang L, Li F, “Learning similarity with cosine similarity ensemble,” Information Sciences, vol.307,pp.39-52,2015.

[27] Hou X N, Ding S H, Ma L Z, et al. “Similarity metric learning for face verification using sigmoid decision function, ”Visual Computer, pp.1-12,2015.

[28] Lin Z C, Wu D W, Hong G E, “Combination of improved cosine similarity and patent attribution probability method to judge the attribution of related patents of hydrolysis substrate fabrication process, ”Advanced Engineering Informatics, vol.30(1),pp.26-38,2016.

[29] Fawaz S. Al-Anzi, Dia AbuZeina, “Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing,” Journal of King Saud University - Computer and Information Sciences.

[30] Nguyen H V, Bai L, “Cosine Similarity Metric Learning for Face Verification”// Computer Vision - ACCV 2010 -, Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers,pp.709-720,2010.

[31] Han Y, Geng Z, Zhu Q, et al, “Energy efficiency analysis method based on fuzzy DEA cross-model for ethylene production systems in chemical industry,” Energy, vol.83,pp.685-695,2015.

[32]Han Y M, Geng Z Q, Liu Q Y, “ Energy Efficiency Evaluation Based on Data Envelopment Analysis Integrated Analytic Hierarchy Process in Ethylene Production,” Chinese Journal of Chemical Engineering, vol.22(11-12),pp.1279-1284,2014.

# 3 研究计划

## 3.1 选题的目的及意义

全国工业接近1/3的能耗源自石油化工行业。作为石油化工的领军行业，乙烯工业的能耗问题研究首当其冲：建立针对乙烯装置的能耗指标体系，并将其应用于乙烯工业，为降低石油化工行业能源消耗树立积极的标杆。能耗指标体系建立在对海量数据的分析之上，这包括：专家经验、操作规范、生产设计数据等。从这些海量信息中挖掘与乙烯能耗息息相关的重要因素，就单/多装置上升至整个乙烯工业，在多个层次框架上建立完整的乙烯能耗指标体系，对企业的乙烯能耗分布加以评估统计，从而寻求节能降耗的路径与指导方案，将量化的降耗目标分解落实。

衡量区域工业发展水平的重要手段之一是考察其乙烯工业的生产能力，因而，提升乙烯装置的经济效益具有极其重要的意义。乙烯装置操作经费半成以上为其能耗所花费，所以建立有效的乙烯装置能效评价指标可以提升装置的竞争能力，为有效节能举措的实施提供极有价值的指导。当前，评价乙烯能效水平的主要依据是Special Energy Consummation（生产每吨乙烯的综合能耗，SEC）。研究表明，乙烯装置能耗会为很多因素所影响。例如原料、负荷变化、规模变化、技术工艺等因素都会造成乙烯装置能耗的改变。这其中，原料组成是最大诱因。然而SEC却并未考虑到原料组成这一因素。为此，有必要寻找到可以更好地综合影响乙烯能效所有关键因素的能效分析与评价指标，建立相应的能效指标与节能潜力的分析体系，提高将评价与分析结论用于指导乙烯装置实际生产运行的价值。

对于能效指标与节能潜力的分析体系，很重要的两个应用是评级与预测。

## 3.2 主要研究内容

关于乙烯能效评级与预测系统的研究，课题的研究内容主要从以下几点展开。

（1）借鉴Fish-Search算法获取相关页面的方法，获取所需数据。将不同数据表看作不同网页，将不同数据表之间的联系看作网页之间的链接关系，选择合适的相似度判断方法，建立向量空间模型，选取适当的相似度算法，并设定合适的阈值，在数据库中对每条数据逐一判别，获取所需数据。

（2）基于第（1）步，我们可以获得用同一生产装置生产的相似的数据，可以通过合适的数据融合方法对输入的生产数据进行能耗指标以及产量的预测；

（3）基于第（1）步，搜索到输入数据所有相似的数据，先对所有可能的影响因素进行概率分布分析，再联合各能效指标进行分析，影响因素作为前件，能效指标作为后件，影响因素的联合概率作为重要度，推理相应的知识规则，来分析各影响因素对能效指标的影响度，从而进行评级。

## 3.3 研究方案

### 3.3.1 技术方案

针对课题研究内容，对应解决的技术方案如下：

（1）针对第一点研究内容，我们在选取所需数据时，先研究传统的基于网页的爬虫算法，深入学习其他学者提出的关于Fish-Search算法改进、相似度算法应用的论文，考虑如何将其改进应用到数据库中。

相似度算法是Fish-Search算法的核心，我们考虑选择使用余弦相似度算法，来评估两组记录之间的相似性。并考虑将余弦相似度算法进行改进，因为虽然余弦相似度对个体间存在的偏见可以进行一定的修正，但是因为只能分辨个体在维之间的差异，没法衡量每个维数值的差异，会导致这样一个情况： 比如用户对内容评分，5分制，X和Y两个用户对两个内容的评分分别为(1,2)和(4,5)，使用余弦相似度得出的结果是0.98，两者极为相似，但从评分上看X似乎不喜欢这2个内容，而Y比较喜欢，余弦相似度对数值的不敏感导致了结果的误差，需要修正这种不合理性，就出现了调整余弦相似度，即所有维度上 的数值都减去一个均值，比如X和Y的评分均值都是3，那么调整后为(-2,-1)和(1,2)，再用余弦相似度计算，得到-0.8，相似度为负值并且差异不小，但显然更加符合现实。在能效指标库中，每一维的数据相差都不会很大，所以我们借鉴这个修正方法来改进余弦相似度算法，结合到鱼群算法使用。

（2）深入学习多维概率分布的相关知识，透彻学习其他评级方法。同时学习关于知识表示规则的方法，重点学习模糊规则表达方法。研究现有的多维概率分布或二维概率分布的应用，考虑将二维概率分布拓展到多维。学习关于Moran方法，Box-Cox 变换和多项式正态变换的文献，将其应用到对乙烯能效各种可能的影响因素的概率分布上，根据概率建立知识的规则库，然后对其进行评级。

（3）学习关于数据融合算法的相关文献，选取合适的算法对输入的数据进行产量及能耗指标的预测。

### 3.3.2 实施方案所需要的条件

课题的研究需要大量的文献信息作为基础，因此前期的工作需要就是阅读课题相关的国内外文献，了解基础算法并对相关算法有清晰的分析，充分利用网络资源、学校图书馆各项资源，最后确定课题的详细方法；后期主要工作需要就是课题算法的具体编程以及收集数据进行实验验证分析，因此只需要可用于编程环境的计算机一台以及实验验证所需要的数据集。

### 3.3.3 拟解决的关键问题

课题的关键问题有三个：

（1）其一是对爬虫算法的改进，由于指标数据，能耗数据及生产数据不在同一个数据表中，需要找到在数据库的多个表之间建立关联的方法，然后用修正的余弦相似度算法对输入的数据做相似度分析，找到相似度满足要求的数据集；

（2）其二是评级系统，有两个关键点，一是评级指标的选择，对各项可能的影响因素做概率分布；另一个是评级方法的选择，考虑选择概率分布方法，将影响因素联合能耗指标，如SEC、EPI等，作概率分布，分析各影响因素对能耗指标的影响概率，根据联合概率分布，可以用知识的模糊规则表示方法进行分析，评出等级。

（3）第三个是基于爬虫算法得到的数据进行预测，应选择合适的数据融合算法，对只输入生产数值的数据进行产量及指标的预测。

## 3.4 课题难点分析

（1）对于课题研究内容的第一点，难点在于如何在根据一条数据做相似度计算找到多个表中的对应数据，可以考虑先从总表（包含生产数据，各单项能耗，综合能耗）中找到相似度满足要求的数据，再根据这些数据，根据时间、生产技术、生产装置等等，找到对应的各项指标，整合成一条数据；除此以外，还应选择合适的相似度阈值，过高的话会使留下的数据太少，后面进行预测或评级时数据量太少，意义不大；若阈值过低，则会使数据不具有代表性。

（2）对于第二点，利用概率分布确定各影响因素对能效指标的影响，难点在于确定多维概率分布的概率密度函数。确定好多维概率密度函数之后，根据概率密度和评级指标建立的规则也是难点之一。除此之外，评级指标和评级方法的选择都是需要不断试验来确定的。

（3）在利用改进的Fish-Search算法和合适的数据融合方法进行产量和能效的预测时，如何选择合适的数据融合方法是难点。

## 3.5 预期研究成果及创新点

预期的研究成果主要有五方面：

（1）第一，通过实验研究分析，改进Fish-Search算法，并能将此算法成功应用于数据库多个表中，选择合适的相似度判断方法，筛选出所需要的数据；

（2）第二，能够选择出合适的概率密度函数，将多个影响因素作为多维的随机变量，计算出其概率分布，并能联合能耗指标，如SEC、EPI等，作概率分布，根据联合概率分布，针对概率分布做出知识表示规则，分析各影响因素对能耗指标的影响概率，评出等级；

（3）第三，利用合适的数据融合算法，基于利用改进的Fish-Search算法所得到的同一装置的数据，对只输入生产数值的数据进行产量和能耗的预测。

（4）第四，用乙烯能效生产数据测试此乙烯能效评级与预测系统，验证此系统的实用性和算法的正确性。

（5）在研究的工作基础上，撰写硕士论文，发表学术文章1-2篇。

创新点：

（1）将Fish-Search算法进行改进，应用到数据库中，找到所需数据集；

（2）基于多维概率密度分布理念，通过不断的试验，做不同的影响因素概率分布，获取对能效指标影响比较大的因素；

（3）对能效指标及其影响因素做联合概率分布，并根据知识的模糊规则表示方法，对其进行评级；

（4）基于改进的Fish-Search算法和恰当的数据融合算法，对乙烯生产数据进行能耗和产量的预测。

## 3.6 工作计划进度

2016年07月 ~ 2016年10月 整理相关文献，明确课题，建立指标模型库

2016年11月 ~ 2016年12月 实验方法设计

2016年12月 ~ 2017年02月 实现算法并应用于数据集验证算法的有效性

2017年03月 ~ 2017年05月 对实验结果分析整合

2017年06月 ~ 2017年11月 根据实验结果进一步深入研究

2017年12月 ~ 2018年04月 撰写大论文

|  |
| --- |
| 指导教师意见：  指导教师签名：  年 月 日 |
| 审核小组意见：  审核小组组长签字：  年 月 日 |
| 研究生根据审核小组意见对开题报告的改进措施：  年 月 日 |
| 备注： |