南京大学

研 究 生 毕 业 论 文

（申请工程硕士学位）

论 文 题 目	基于关键词聚簇技术的广告检索系统设计与实现
作 者 姓 名	顾 进 杰
学科、专业名称	工 程 硕 士（软 件 工 程 领 域）
研 究 方 向	软件工程
指 导 教 师	郑滔 教授　刘钦 讲师

2011 年 5 月 13 日

学　　号：**MF0932016**

论文答辩日期：　　年　月　日

指 导 教 师：　　（签字）

基于关键词聚簇技术
的广告检索系统设计与实现

作　　者：　　顾进杰

指导教师：　　郑滔　教授　刘钦 讲师

南京大学研究生毕业论文

(申请工程硕士学位)

南京大学软件学院

2011 年 5 月

The Design and Implementation of Advertising Retrieval System Based on keyword Cluster

GU, Jinjie

Submitted in partial fulfillment of the requirements for the degree of Master of Engineering

Supervised by

Professor **ZHENG, Tao**

Lecturer **LIU, Qin**

Software Institute

NANJING UNIVERSITY

Nanjing, China

May, 2011

摘　要

内容页广告，即通过用户浏览的页面内容以及用户浏览的历史记录(cookie)，有针对性的进行广告推送的一项互联网广告技术。目前内容页广告平台在国内蓬勃发展，是许多网站赖以生存的变现手段。在内容页广告系统中，存在两种基本的广告类型：相关性广告和站点匹配广告，本文主要关注的是基于相关性广告的一种新广告投放方式。

文章的主要目的是在内容页广告的检索端引入了一种轻量级模型（关键词聚簇）来解决在目前内容页广告检索结果中相关性广告缺乏的问题。该技术主要是基于页面的关键词特征，通过关键词聚簇的方法，对特征进行扩展，引入了更多语义以及意图上匹配的广告。通过这种方式，进一步扩大广告系统的相关性广告检索能力，为用户提供更好的服务。

文章的主要工作就是将该技术引入到内容页广告的检索端。首先，本文介绍了目前在内容页广告中流行的技术，包括通用的主题模型以及关键词扩展等等，同时引出了关键词聚簇技术，并分析了该技术在广告检索中的作用和优势。在后续章节，通过整个系统的需求以及检索端的流程来详细介绍该技术。在这个过程中引入了对自反馈体系的解读，进一步体现了该模型轻量级、自完善的特点。接着着重介绍了该技术的概念以及整体的设计，根据目前检索端的架构和模块划分，综合了模型的特点和系统的框架，着重展开介绍如何设计和实现，最终使得该检索流程在现有系统上稳定的运行。在这过程中还讨论并分析了系统中出现的一些性能以及模型上的问题，提供了更好的解决方案，并且通过具体的性能数据来说明解决方案的提升点。最后，详细分析了关键词聚簇技术对整个内容页广告的检索端带来效果上的提升。

内容页广告，主要通过相关性来衡量广告和页面的关系。相关性广告的资源扩大之后能够给平台效果带来较大的提升。通过本文对关键词聚簇技术在检索端应用的研究，解决了应用和实现上的一些难题。进一步的增大了相关性广告在系统中的占比，并且证明了该途径能够提升广告平台的效率和业务数据。文章的最后部分，通过总结与展望，对该技术以及应用前景做了一些分析。

关键词：内容页广告，广告相关性，关键词聚簇，自反馈

Abstract

Contextual advertising, which is also called content match advertising, refers to the placement of advertisement on web pages or content displayed in mobile browsers. At present the content match advertising platform is booming, and many web sites are depending on this platform to make money. In the content match advertising system, there are two basic advertising types: relevance ads and site targeting ads, this paper mainly concern is a new advertising mode which is based on the relevance ads.

The main purpose of this article is to introduce a lightweight model(keywords cluster) to solve the lack of relevance ads in the search results. This technique is mainly based on the page keywords, through the keyword cluster algorithm, to expand the keywords list. In this way, system will target more semantics matching ads or intentions matching ads, and the advertising system can provide users with better service.

The main work of the article is to introduce this technology into the retrieval model. Firstly, this paper introduces the present technologies in the content page ads, including the topic model and keywords expansion model etc. Meanwhile, the paper introduces the keywords cluster technology, and analyzes the advantages of this technology in advertising retrieval model. in subsequent chapters, introduced the needs of the technology and the auto-feedback system. The paper also introduces the concept of keyword cluster and overall design of this technology. The paper also discussed and analyzed the runtime performance of the technology and the emergence of the model. The author provides a better solution to solve these problems. Finally, we analyzed the effect of the improvement after the system using keyword cluster in this system.

Content matching ads, mainly through relevance score to calculate the relationship between the ads and the page. The expansion of the relevance ad resources will bring greater effectiveness for the platform. The research in this paper solve some problems in the application and realization of this technology, further increased the proportion of relevance ads in the system, and showed that keyword cluster model can improve the efficiency of the advertising platform. The last part of the paper, we analyzed the prospect of this technology through the summary and outlook.

Keywords：contextual advertising, relevance of advertising, keyword cluster, automatic feedback.

目　录

图目录

表目录

第一章 绪论

1.1 选题的背景与意义

内容页广告系统，即通过用户浏览的页面内容以及用户浏览的历史记录(cookie)，有针对性的进行广告推送的技术。该题是在百度的内容页广告系统（百度网盟）的检索端实习过程中，参与设计与研发的一个检索端策略。

内容页广告，顾名思义，就是需要展现在具有一定的主题含义的网页中，展现的广告。但在实际的互联网中的流量，不完全是内容页流量。举个例子：小说页面、视频网站、音乐站点，就不是内容页面。本文关注的技术主要是应用在内容页面上，那内容页究竟有什么样的优势呢？这需要从互联网广告说起。

当今的互联网，主要的三大课题就是：移动、广告和云。可见广告在互联网生态圈中无与伦比的重要性，当 pc、客户端、操作系统开始淡出人们的视线之后，广告依旧坚挺的站在互联网浪潮的尖峰上。互联网广告的优势可以归结为以下几点：

1. 丰富的媒体资源，呈现形式多样：由于电视广告的到达不精准性，所以广告的投放容易出现集中扎堆央视或者卫视的情况，而有些小型企业则因为电视广告的费用高的原因而望而却步。网络广告则因其精准定向性而避免了这些现象的发生。精准定向让广告只投放给相应的目标受众，每个广告都只会出现在与其产品定位适合的网站媒体。互联网中，媒体流量丰富，广告的受众面大。

图 1.1 互联网广告的优势

2. 精准的定向技术，更加贴近用户：互联网广告精准到达可以从三个维度来实现。时空维度上，通过每天不同品牌产品、不同人群的浏览峰值得出最佳投放点；在外围环境方面，甚至可以根据天气的实时变化进行广告素材的差异化调整，投放匹配用户周边的环境，投放增加广告的亲和力；根据媒体所说的内容实现兴趣定向，广告出现于媒体内容相关，根据页面内容出现相匹配的广告，捕捉用户最高关注点峰值。
3. 实时的优化过程，降低成本：互联网广告在投放过程中，通过对独立 Cookie 的展现频次的控制，做到对用户的实时连续跟踪，提高广告的精准到达。在一般意义的广告中，通常都是按照展现或者位置进行付费。比如电视媒体，黄金时间的广告价格非常高，但是如何衡量广告的转化效果则比较难。但是在互联网广告中，按照点击进行计价，这样无形之中，节省了广告主的费用。[李东，2005]

而互联网广告的这三点优势，内容页广告都具备。内容页广告，是互联网广告中更加特殊的一部分。这些网页都含有比较明显的主题，广告平台的检索端通过分析用户浏览页面的内容信息，提取页面的关键部分，根据这些信息来触发广告。目的是精确匹配广告和浏览者，尽最大的可能命中目标人群。内容页广告吸收了互联网广告的精髓，通过丰富的媒体展现形式包括图片、文本、flash、悬浮广告等等，按照点击计费，平均的展现价格（acp）低，投放的精确度较高，能够为广告主带来更多的收益，是互联网广告中非常重要的一个组成部分[朱彤，2010]。

但是要想做好内容页广告，却非常难。列举几点：

1. 精确匹配的难度太大：互联网广告精确的要求，本质来说是浏览者得意图与广告的匹配。但是用户的意图，是一件非常虚幻的东西。尤其在互联网广告中，通过网页这样的媒介，来映射用户的意图，但实际上，却很难表达。例如，用户在看汽车配件的网站，从页面的分析来看，应该是与汽车配件相关，可是用户究竟是想买汽车？还是想修汽车？还是想改装汽车？检索端只能无从下手了。
2. 优质流量稀缺：内容页广告，从字面分析，就对页面有一定的要求，需

要能够根据页面分析出一定的主题特征。但是在实际的互联网中，真正符合内容页标准的，少之甚少。大量的娱乐流量、视频网站占据着主要的互联网流量（除了搜索引擎流量之外）。

3. 广告用户体验：内容页广告，牵涉到很多不同的用户和对象：网站主、广告主、检索技术提供商（百度、google 等）、网页浏览用户。而用户体验是否好坏，则直接影响了这生态链。广告说到底都是被用户浏览的，如果广告的用户体验非常差，比如会有巨大的噪声、闪动、不良信息，会给网站带来负面的影响。
4. 广告主的 ROI 究竟如何衡量：虽然内容页广告，以点击广告计价，有比较低廉的展现价格（acp）。但是广告主既然在上面投入了资金，那就需要能够看到一定的回报。但是在互联网的商业中，广告的被点击行为离广告主最终的要求还差很远。有些广告主希望用户能够发生购买行为，有些广告主希望用户能够发生注册、发表评论，而有些则希望用户在网站上留下一些信息。而这些，都不是能够通过点击这个行为来简单的衡量的。这也让检索端的优化[ccw，2011]非常的艰难。

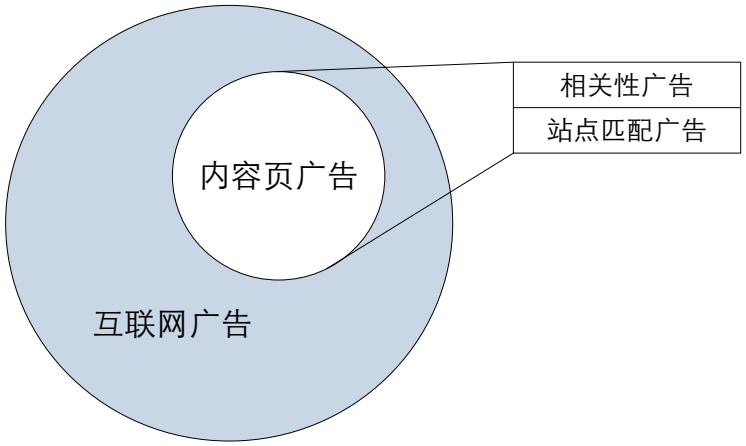

图 1.2 互联网广告与内容页广告的关系

在内容也广告中，也有两大主要的广告类型：相关性匹配广告与站点匹配广告。相关性匹配广告主要的手段是通过广告主选择的题词与用户浏览的页面的特征进行匹配，进行广告的展现。站点匹配广告主要是通过广告主选择定向投放网站，主要是发挥广告主本身的意愿。通过业务数据以及用户反馈数据比较可以发现，关键词匹配广告能够在内容页面上给广告带来比较好的展现效果以及 ROI(Return On Investment)。

但是在百度网盟的广告体系中，关键词匹配广告与站点匹配两者的流量比率有所失调，站点匹配广告的流量占比比预期高。造成这样现象的原因与系统的广告的资源有关，在广告库中，选择关键词匹配投放的广告主不够多，因为这样的投放方式需要广告主更多的主动性。在有限的资源中，又由于检索方式比较单调，大量的广告很难有机会展现。站点投放广告居多，这样不能给整体带来比较好的转化效果。为了提升整个系统的转化效果，我们需要在原有的关键词广告算法中，加入新的血液。

关键词聚簇的方法，是通过 offline 的挖掘方式，通过不断的学习，异步的更新 offline 知识库。在线上系统中，通过对页面提词进行图扩展，通过这种方式，进行广告的触发，为关键词广告带来更大的展现机会，进一步的提高系统的业务数据与广告主的 ROI。其主要的手段，就是通过扩大关键词匹配广告的数量，我们称之为广告召回。举个例子：如果通过页面特征分析模块，我们得到了页面的关键词特征（term）是“轿车”，在原先的触发过程中，我们必须要求广告的拍卖词中含有“轿车”字样，广告才有可能触发。但是不含有“轿车”的广告就不应该被触发吗？当然不是，比如“宝马”、“奔驰”都可以认为是与“轿车”意图相关的词语，喜欢“轿车”的用户，理论上也是与“宝马”意图相关的。关键词聚簇，本质上就是希望能够找到“轿车”→“宝马”这样的意图映射，通过在检索端的发力，来进一步提升关键词广告的占比，为广告主带来更好的转化效果。

1.2 国内外研究的现状及分析

目前在国内外基于搜索以及内容页的广告服务非常多，比如 google AdWords/AdSense，Yahoo!Advertising Solution[google，2011]等等。在国内，这样的广告服务也非常多，比如百度、搜搜、搜狗、有道等搜索引擎都提供搜索广告的服务，目前国内最大的搜索推广是百度的凤巢。在国内的内容页广告推广服务，主要还是 google AdSense[Harold，2006]和百度的网盟[baidu，2011]推广两家。

内容页广告推广，主要是通过一种联盟网站的形式，在非自有流量上进行变现的手段。目前的主流方式有两种，一种是相关性匹配广告，一种是站点匹配广告。两种方式都是通过定向方式帮助客户锁定目标人群，并以丰富的样式将客户

的推广信息展现在目标人群浏览的各类网页上，在其上网全程产生深入持久的影响，有效提升客户的销售额和品牌知名度。并且通过按效果付费的方式，进行转化[sina，2007]。

本文主要关注的是相关性广告投放方式。在内容页广告系统中，目前主要的相关性投放手段是通过关键词匹配，需要通过页面特征模块进行所有网页及站点信息的抓取，并且通过线下的算法进行页面特征的提取。页面的特征分为很多类，主要的特征是页面的关键词(term)特征。同时，广告系统中维护所有的广告特征，对每一条含有主题词的广告，进行关键词的特征提取，建立检索倒排拉链。触发的过程，就是通过页面的 term 特征，在广告库中进行倒排检索，并且根据检索的中间信息，进行广告相关度的排序(rank)，最终通过计价并最终展现在用户页面上。

只依靠页面的题词，触发召回（触发得到的广告候选集数量）是远远不够的。如果召回不够，则关键词广告的展现占比就会比较少。如果关键词广告的展现占比不够，则在与站点投放广告的较量中就会处于下风，往往会带来系统的整体收益下降，广告主的 ROI 不高等问题。

目前国内在解决该问题上，有比较多的方法和手段。

例如使用主题模型，比如 PLSA(Probabilistic Latent Semantic Analysis) [Wiki-PLSA，2011]或 LDA(latent dirichlet allocation) [Wiki-LDA，2011]。 LSA 是处理主体模型中的著名技术。其主要思想就是映射高维向量到潜在语义空间，使其降维。LSA 的目标就是要寻找到能够很好解决实体间词法和语义关系的数据映射。PLSA 是以统计学的角度来看待 LSA，相比于标准的 LSA，他的概率学变种有着更巨大的影响。概率潜在语义分析基于双模式和共现的数据分析方法延伸的经典的统计学方法。概率潜在语义分析应用于信息检索，过滤，自然语言处理，文本的机器学习或者其他相关领域。概率潜在语义分析与标准潜在语义分析的不同是，标准潜在语义分析是以共现表（就是共现的矩阵）的奇异值分解的形式表现的，而概率潜在语义分析却是基于派生自 LCM 的混合矩阵分解。考虑到 word 和 doc 共现形式，概率潜在语义分析基于多项式分布和条件分布的混合来建模共现的概率。所谓共现其实就是 W 和 D 的一个矩阵，所谓双模式就是在 W 和 D 上同时进行考虑[Thomas，1999]。

在应用中，搜索引擎根据页面的关键词特征，进行主题模型的计算。计算出该关键词簇所属于的主题类别。并且根据 top N 的主题类别的支撑词进行触发关键词的扩充。通过这种方法，能够很好的进行触发词源的扩展处理。

关键词聚簇技术同主题模型在内容页检索上的应用相比较，具有更加灵活、轻便，迭代速度更快的优点。主题模型往往对机器资源、训练语料的要求更加严格。关键词聚簇技术往往能够单机运行，并且能够灵活使用各种有价值的资源，加快自身的迭代。

除了主题模型，目前还有比较流行的算法是直接进行关键词集合的扩大，比如使用 wordsim 关键词扩展。通过对搜索 query 的检索结果摘要进行挖掘分析，给出 term 之间的 sim 值，对 sim 值较高的 term 对进行关键词的扩展并且进行触发。

在进行广告相关度计算的时候，传统的方法使用 KL 的方式来进行。KL 的定义是给定真实的概率分布 P，以及另外一个用于估计 P 的概率分布 Q，KL-分散度定义为下列公式（应用中取负值）。在广告检索里，假设真实分布是查询 R 的相关性模型（request），近似分布为文档语言模型（ad），这个没有疑问。但是在 KL 的公式中，只有一部分依赖于文档，故如果只使用 request 原词，可以用简单的极大似然估计来推导公式中的两个概率；由于广告匹配中单 term 匹配常见，通常会引入扩展词项，来平滑查询和相似文档。在相关性模型中，P（w|D）仍使用极大似然，因此问题转化为如何去计算 P（w|R）。目前大部分使用的 KL 的方法和关键词聚簇有一定的不同，将两者做一个比较：

都为扩展技术提供一个正式的检索框架

KL 强调查询扩展（网页端），关键词聚簇（cluster）强调文档扩展（拍卖词端），但这点不是本质上的差异。

从排序公式上来看：

关键词聚簇公式中 ： score= $\sum_{t}^{n} P(t|request)P(t|ad)$ 其中 P（t|ad）主要有两个因素：

1）节点内的重要性权值

2）节点之间的关联权值（可看作相似文档的相关度）

KL 公式中 ： score= $\sum_{t}^{n} P(w|R)\log P(w|D)$ 其中，P（w|R）主要也有两个

因素：

1）查询的相似文档中，对 w 的语言概率的权值

2）查询-相似文档的似然得分

关键词聚簇除了在进行广告索引中能起到一定的作用，在广告的相关度计算上，也能很好的利用本身模型的特点，进行广告相关度的计算。KL 距离和 cluster score 的计算，都能够很好的解决相关性计算的问题，两者相比，KC 在更加贴近广告拍卖词的情况下，通过类似主题方式的计算，能够更好的符合广告相关性的要求。

该技术同这种传统的关键词扩展相比较，具有更好的灵活性和准确性。关键词扩展，往往会引入比较多的噪音，并且不可控。而使用聚簇的方式，通过权重的控制，能够很好的控制扩展的质量以及扩展的数量。

1.3 课题来源及主要研究内容

课题来源于本人在百度商务搜索部（ECOM）网盟产品线的实习经历。百度的网盟推广能够帮助中小网站获得广告收入，同时帮助广告主进行广告的精准投放，是百度在非自有流量上变现的重要产品线。百度将互联网众多内容网站整合，建立了国内最具实力的联盟体系；百度联盟囊括了 24 个行业类别的优质网站，加盟合作网站累计超过 30 万家，影响力覆盖 95%以上的中国网民。主要的优势在平台大，依托百度搜索巨大的流量和数据支持，客户数量庞大，在中国的互联网广告业务中占据重要位置。

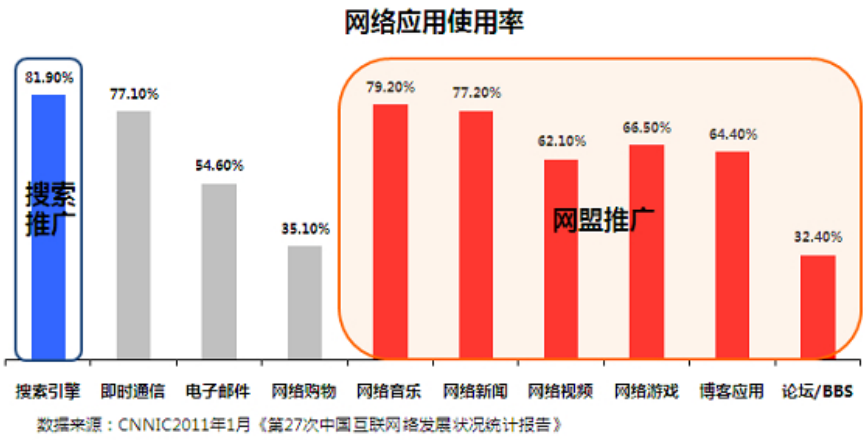

图 1.3 网盟投放流量的优势

该课题的创意来自于网盟检索端的需求。网盟的检索端，所处理的主要的工作，就是根据网民的点击，根据不同的流量展现不同的广告。网盟中的流量很大，每天的检索量在几十亿的级别，流量分成内容页和非内容页，比率大概为 3:7，内容页的流量虽然少，但是这部分流量客户转化率高、同时 CTR/CPM 较高（注：在互联网广告中非常重要的两个指标：CTR:Click Through Rate-千次展现点击；CPM:Cost Per Mille-千次展现收入，是主要评价检索端效果的两个指标）。

目前的检索端，主要根据网页分析模块产出的触发 term 进行广告的检索。使用 term 进行检索的好处是比较精准，如果 term 的质量较高，则对业务数据的提升非常明显。但是仅仅使用 term 触发，往往不能达到推广的要求，原因在于仅仅使用词，面太窄，不具备主题性质的推广。在网盟的检索端，也存在主题形式的触发，根据题词信息，使用 plsa/lda 模型，进行主题计算，然后再根据主题进行广告的检索。但是主题模型在内容页广告的检索上，存在比较多的问题。问题一：主题模型是一个很通用的模型，通用但并不代表就能很好的应用在内容页广告上，比如在内容页广告中，可以认为所有的股票都是一个 topic，所有的高档汽车都是一个 topic，但是不同的病都必须是不同的 topic，这一点主题模型很难处理；问题二：主题模型需要进行大量语料方面的运算，消耗大量的机器性能，需要考虑并行计算、分布式计算等等资源来训练。

关键词聚簇的方法，是通过 offline 的挖掘方式，通过不断的学习，异步的更新 offline 知识库。在线上系统中，通过对页面提词进行图扩展，通过这种方式，进行广告的触发，为关键词广告带来更大的展现机会，进一步的提高系统的业务数据与广告主的 ROI。它具备训练速度快，规模小，应用性强的特点。相对于主题模型，关键词聚簇的方法，能够加入更多的强关联的关系，例如上述举例的股票医疗的例子，可以方便的解决。另外它不需要进行大量的运算，通过大检索以及 NLP 获取的数据，通过线上的反馈系统，能够获取很好的效果。如何将关键词聚簇的手段运用到检索端，是本文的最大课题。

1.4 本文的工作

本文的课题在百度网盟投放的大平台下展开。在国内外，对语义检索和意图检索都有了比较深入的研究。但是在中国的互联网广告中运用的还不多。同时，

在该投放平台中，也正需要类似的检索方式来进一步扩大内容页广告的投放量。

在这样的背景下，文章的目的就是希望能够将该技术引入目前的检索端，并且通过最终的业务数据以及运行数据，来说明该技术的可行性以及技术优势。文章首先介绍了目前在内容页广告中流行的技术，并在后续章节，通过整个系统的需求以及检索端的流程来详细介绍引入该技术的需求。并在这个过程中引入了自反馈体系的解读，进一步体现了该模型轻量级、自完善的特点。中间部分着重介绍了该技术的概念以及整体的设计，根据目前检索端架构和模块划分，综合了模型的特点和系统的框架，着重展开如何设计和实现，最终使得该检索流程在现有系统上稳定的运行。在这过程中还讨论并分析了系统中出现的一些性能以及模型上的问题，提供了更好的解决方案。

最终，通过一系列的反馈数据，可以较明显的体现出该技术的引入，提升了内容页广告平台的检索能力，并最终为平台带来了收益。

1.5 本文结构

互联网广告在行业中的发展已经有些时日，但是在内容页广告中，还是有一定的发展空间。正文主要是说明关键词聚簇是如何应用在内容页广告的检索端，对百度网盟内容页广告带来收益的：

第一章 绪论部分 主要简单介绍了内容页广告在行业内的发展，以及文本内容的简单介绍。

第二章 内容页广告相关技术概述 介绍与内容页广告相关的技术以及相关评估指标。

第三章 基于关键词聚簇技术的需求 关键词技术如何应用在检索端，这章主要介绍了系统级别的需求。

第四章 检索端的总体设计 网盟中检索端的总体设计，以及关键词聚簇如何嵌入在这样的系统中进行很好的工作。

第五章 模型应用的实现 将关键词聚簇技术应用到检索端后，如果调整效果以及遇到的一些问题和如何解决这些问题。

第六章 总结与展望 对论文期间的工作做总结，以及对关键词聚簇技术在互联网广告中的未来发展做一个展望。

第二章 相关技术概述

2.1 评价广告的相关指标简介

互联网广告由于具有技术上的优势，在效果评估方面显示出了传统广告所无法比拟的优势和特点，具体表现为：

1. 及时性：网络的交互性使广告受众可以直接在线提交反馈意见，广告主可以在几分钟，最多几小时之内收到反馈以了解广告的传播效果、社会效果和经济效果。
2. 方便准确性：网络广告本身具有易衡性，可以方便地准确统计出具体数据，同时，网络的数字化定量分析，可以部分避免在传统广告中因专家意见偏差等主观原因所造成数据失真的情况。因此网络广告效果的调查、评估结果的客观性与准确性大大提高。无论采用何种统计指标，利用软件工具都很容易得到准确结果。
3. 广泛性：网络的广泛性使网络广告效果调查能在网上大面积展开，对极其广泛的调查目标群体进行调查，使参与调查的样本数量增加，范围扩大，有助于提高广告效果评估的客观性和可信度。
4. 客观性：网络广告效果评估无需调查人员出面参与，广告受众不受调查人员的主观影响、不受干扰地回答调查表单上的问题，因此能准确地反映广告受众的态度与看法。故调查结果会更符合消费者的真实感受，更具可信性。
5. 经济性：网络广告效果评估依靠技术手段，与传统广告评估相比，耗费的人力物力少，故成本较低，这也是网络广告效果评估的最大优势。

根据这些特点，对互联网广告的评价以及相关性的一些指标与其他的广告投放类型就大不一样了：

1. 点击率指标（CTR）：点击率是指网上广告被点击的次数与被显示次数之比。它一直都是网络广告最直接、最有说服力的评估指标之一。点击行为表示那些准备购买产品的消费者对产品感兴趣的程度，因为点击广告者很可能是那些受广告影响而形成购买决策的客户，或者，是对广告中

的产品或服务感兴趣的潜在客户，也就是说是高潜在价值的客户，如果准确识别出这些客户，并针对他们进行有效的定向广告和推广活动，可以对业务开展有很大的帮助。

2. 展现收入指标（CPM）：点击率标明了广告站先后被点击的情况，但是根据点击的计价（目前互联网广告中普遍采用 GSP：generalized second price 进行计价）之后，才能得到对网站主或者联盟收入的指标，该指标对衡量产品线的变现能力有一定的指导性作用。可以认为 CPM=CTR* 平均点击价格。
3. 平均点击价格指标（CPC：cost per click）：互联网广告目前主要还是按照点击进行计价，在评价 CPM 以及转化效果的时候，平均点击价格都是一个非常重要的指标，投入回报比，投入则是平均点击价格。
4. 转化率指标（ROI：Return On Investment）：转化率指标是衡量一个广告平台效果的重要指标。但是在互联网广告中，转化很难衡量。一般网络上，“转化”被定义为受网络广告影响而形成的购买、注册或者信息需求。有时，尽管顾客没有点击广告，但仍会受到网络广告的影响而在其后购买商品。转化率在全局上看是一个很难衡量的指标，但是对每一个广告主而言，都有自身衡量 ROI 的手段。
5. 广告相关性：互联网广告中，ROI 很难衡量，那么 ROI 就很难在检索策略的开发中起指导作用，往往只能作为一个后验的评估指标。所以在内容页广告上，应运而生的一个指标就是广告相关性，通过评估广告和页面的相关程度，来指导广告检索策略的开发。

2.2 广告触发概述

广告触发，就是指广告被特征检索出来的过程，本文主要的应用场景就是内容页广告的触发。在网盟的架构中，广告触发可以使用很多的特征。目前主要的特征形式是 term 和 topic id。

在广告的触发中，有两个关键的指标：召回数量和准确率

召回数量：是指一次广告的请求，检索端所能提供的最多的广告候选集的数量。该数值，能够反映出触发资源的丰富程度，也能反映出广告平台的广告资源

的数量。是一种广告检索策略非常重要的触发指标。

准确率：在广告召回的基础之上，检索端需要进行更加细粒度的初选和优选，才能最终获得的符合相关性、有效的广告，这部分广告同召回广告的数量相比，得到的数值为一次广告触发的准确率。

召回和准确是在检索端衡量一个触发策略重要的依据。后续会继续讨论关于这两个指标。

2.3 相关性评分模式简介

在内容页广告中，开发了一个新的策略或者新的检索方式，需要通过相关性的评分来对这种更新打分。目前的打分方式是使用分档的模式，具体的标准可以分为 5 档[于奎，2004]：

1 档：页面与广告的主题完全一致，包括地域、人群，如宝马-宝马

2 档：页面与广告强替代性需求，但主题完全一致，如宝马-奔驰

3 档：页面与广告弱替代性需求，主题基本一致，如宝马-QQ, 或人群相关 手机—笔记本

4 档：主题相关的 B2B 或主题不相关的 B2C

5 档：主题不相关的 B2B

2.4 关键词聚簇技术简介

2.4.1 关键词聚簇

关键词聚簇技术，主要的技术点就是基于 query 的关键词聚簇的图数据。为了获得这样的聚簇数据，首先，需要根据多种特征，将 query 融合为簇结构，要保证簇内元素关注的特征一致，例如在下图中的“白癜风”和“白驳风”。获得聚簇结构之后，需要进行类别隶属信息、二元关联、向量扩展等一系列操作，最终获得图数据结果。

这里主要的工作分两部分，如何获取簇数据和如何获取以及计算有向边关系。

关键词聚簇节点：cluster 类型，将 query 降维。簇数据具有多维的属性信息，购买客户数、pv、地域、产品、类别隶属、TAG、扩展向量信息、统计信息等应

用相关的特征。所有的簇中的元素，在这些属性上要保持基本的一致，例如根据统计挖掘得到“java 软件培训”和“.net 培训”具有比较强的购买关系，在广告主的购买关系中非常明显，他们具有比较强的类别特性（计算机培训类别）、统计数据上共现次数很多，他们可以成为一个簇节点。在网盟的检索端，会从大检索的历史 query、网盟的广告主购买关系等数据结果中进行挖掘这样的相似结构。再根据 query 中文表达结构的特征，多种方法融合，得到的候选集合经过语义、用户行为的校验，能够大大提高精确度，最终形成所需的簇节点。

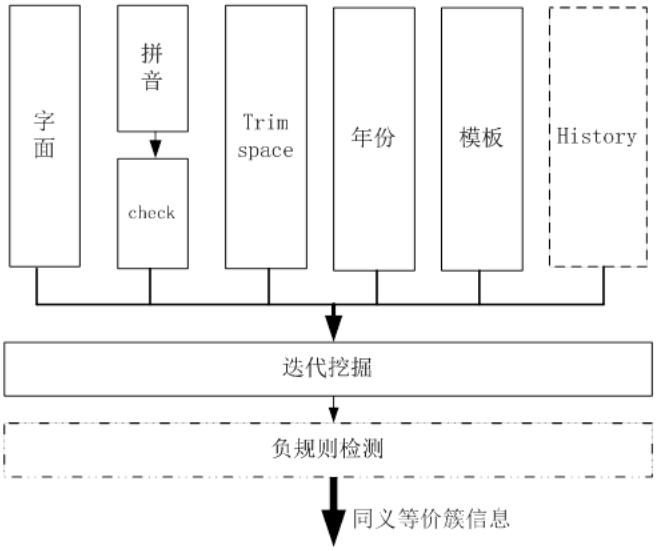

图 2.1 簇节点数据迭代挖掘

有向边：定义为“二元推荐关系”，可以理解为 A→B 这样的模式，其中 B 是具有一定商业价值的词，举个例子：“张学友”→“刘德华”是个比较符合意图的二元关系，但是由于“刘德华”并不具备商业价值，所以这不是一个好的二元关系。但是“张学友”→“演唱会门票”是一个很好的二元关系的示例。二元关系在特定的空间共现是挖掘有向边候选的重要手段。

目前我会使用到的一些挖掘算法有：

1. 广告商购买行为关联挖掘：广告商购买行为很好理解，搜集联盟历史上所有的广告商关键词的购买历史，但是有一定的局限，仅适用于拍卖词内部，并且只适用于一般的大客户。
2. 基于搜索摘要的扩展挖掘：搜索的摘要内容非常丰富，同时，关键词挖掘主要针对的对象是 query 和拍卖词（广告主参与竞价使用的关键词），可以挖掘到在摘要中拍卖词和 query 共现的特征。

另外还有很多的挖掘方式，比如用户 session 中查询行为的关联挖掘、分层规约等等这里不再赘述。通过在行为上共同出现，经过一些语义的校验，加上意图匹配，最终可以形成具有商业价值的二元推荐关系。这样基于聚簇和二元关系形成的关键词聚簇的图关系就基本形成了。最后，需要通过多种手段的校验流程，对二元关系进行最后的过滤，得到最终的数据形式的关键词簇。

2.4.2 关键词聚簇技术的模型简介

“关键字聚簇”方法相当于一种为广告库定制的语言模型[宗成庆，2008]，它把关键词和检索端的 query 组织成图的形式：

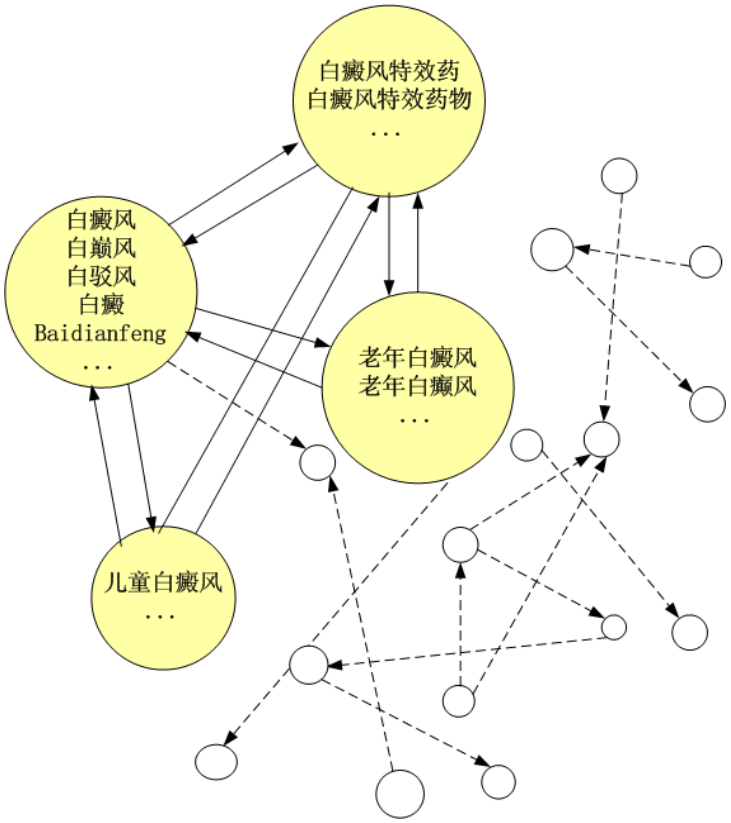

图 2.2 关键词簇图数据

节点是根据关键词或者检索 query 聚类得到，节点内所有的元素元素做到主题相关；边为词之间的关系：广告库的购买，搜索的同现次数，互信息或语义相似，是一个扩展性较强的排序模型。

引入关键词聚簇的技术以后，会需要在现有系统上做如下改进：

1. 相比原先的关键词触发来说，克服了词项独立性假设，加强了相关词概念。

2. 现有的簇及其拓扑关系使得 page-ad 之间存在多条路径，基于图的排序方法可以有效地利用节点之间的依赖关系，得到各个节点的权重, 使用多条路径的产生概率估计相关度。
3. 相比 topic model 来说，它的计算简单，可控性强，可以通过数据优化和裁剪得到匹配性能和匹配质量的提升。

检索的算法设计要考虑到检索效率以及在广告库中索引效率及召回，关于这几点，关键词聚簇算法有一定的优势：

1. 基于簇的触发/扩展效率上高于简单的对词的扩充，准确性相对可控；例如“游戏”“汽车”“股票”只需要一个 cluster cover，索引性能得到优化。
2. 在额外召回的前提下，相关性准入条件能够趋严。相关性排序的截断阈值：
3. 不同 request，广告库的资源不尽相同，现在的截断阈值较低，不区分精准广告和擦边球。在相关广告足够充裕时做动态截断，防止不相关的广告侵占优质流量。

2.4.3 相关性计算的理论模型

关键词聚簇是基于图模型的，但是纯的有向图模型非常复杂，加上关键词节点以及聚簇数量较多，可能会导致检索时候消耗大量的性能。所以在将数据应用到检索端的时候，将关键词聚簇的模型简化：

1. 通过关键词聚类的方法，获得关键词簇
2. 通过对簇内的关键词进行赋权，获得关键词的赋权数据。获得每一个触发关键词同每个簇之间的关系，最终获取所有的关键词到簇的拉链：<term,[cluster]>，在这里基于关键词的赋权方案来进行权重计算，及

$$P(term_i \mid cluster) = \mu(term_i, cluster)$$

3. 同时，由于每一个广告都含有关键词，再根据簇的支撑关键词，获得簇和广告之间的关系：<cluster,[ad]>。如何评价一个广告和簇的关系，这里需要一个权重来进行标示，计算的方法是根据一个关键词在广告所有关键词中的最大似然分布来表示：

$$p(cluster|ad) = \frac{C_{ad}(keyword, cluster)}{C_{ad}(keyword)}$$

4. 最后，在聚簇的内部，需要发掘两个簇之间的关系，根据簇的支撑元素的 overlap，计算出两个簇之间的关联，将这些关联拆散，同时只保留一层扩展，形成簇之间的关系索引：<cluster,[cluster]>

最终计算相关性 score 的时候，将上述的分值进行 combine:

$$\text{RelevanceScore}(\text{page}，\text{ad}) = \text{weight}_{\text{trigger-term}} * \varphi(\text{trigger} - \text{term}, \text{ad})$$

$$= \sum \text{weight}_{\text{trigger-term}} * [(\text{P}(\text{term}|\text{model}) * \text{P}(\text{model}|\text{ad})]$$

$$= \sum \text{weight}_{\text{trigger-term}} * [\sum (\text{P}(\text{term}|\text{cluster}) * \text{P}(\text{cluster}|\text{cluster}') \text{P}(\text{cluster}'|\text{ad})]$$

得到最终的 relevence score，用来进行广告初选的排序分数[苑春法，2005]。

2.4.4 关键词赋权方案设计以及对比

在<term,[cluster]>中需要评估一个 term 和簇的关系，最终给出一个标示该 term 和簇之间关系的权重：P(term|cluster)。这里需要对比不同的赋权方案对数据的准确性带来的影响。

目前在检索端使用的比较多的几种关键词赋权方案有：TF*IDF、基于 term 的分档赋权、plsasim、wordsim 等，关于这几种赋权方案，在关键词聚簇的调研过程中都使用过，对于不同的赋权方案进行了综合的比较：

表 3.1 关键词赋权方案对比

赋权方式	优点	缺点
TF*IDF Term frequency & Inverse document frequency TFIDF 的主要思想是：如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TFIDF 实际上是：TF * IDF，TF 词频(Term Frequency)，IDF 反文档频率 (Inverse Document Frequency)	是一种经典理论模型，计算方式简单	1. general term 容易被加权，比如说“治疗”“医院”“德国” 2. 受限于聚簇算法。考虑到词频和簇的大小之间的 normalize，会造成赋权的不当 3. 区分能力不好（即使*idf），比如一些 case，同这种方法无法进行比较恰当的赋权，导致 badcase，比如： “豆芽”到主题簇“豆芽生产机”， “牛肉”到主题簇“牛肉浸膏” “芝麻”到主题簇“芝麻花岗岩”
基于 term 的分档赋权 即通过 term 的商业价值字典，将 term 分成几个 level，对不同 level 的 term 赋予不同的值	分档赋权，去除了很多的赋权错误，并且在赋值上具有有一定的区分度	1. term 赋权区分度不够，好坏的差距不够明显(分档)，即使在权重第 1 档的 term 中，也会出现一些比较差的 case 2. 倾向于 bidnum（手机、汽车、中控、扑克)，但实际上这些词在页面中只起到修饰作用（手机导航软件下载、汽车贴膜、中控防盗门、扑克透视眼睛），造成一些赋权的错误
Plsasim 通过 plsa 模型，对 term 赋	通过 plsa 校验 term weight 有	1. 区分度不够明显，尤其是对要求比较精准的部分，在准确率上不

权进行指导	一定的区分度	够。 2. plsasim 单独不足以做决策，分值的分布比较平均
Wordsim 主要获取数据的方式是通过检索端的摘要，线下训练获取的词表	通过权重截断的调节，能够在精度和召回中进行权衡 区分度明显	1. 在中低档的数量较多，需要进行一个比较好的权值调整 2. 目前使用 term 赋权，就采用了 wordsim 的方式

最终的赋权方案，综合以前的优缺点，目前使用的是 wordsim 版本的赋权，通过大检索的数据，能够更加贴近网友的习惯。

2.5 本章小节

内容页广告是一项需要综合考虑多方面因素、需要进行多项指标的权衡、复杂的系统，在这样的系统中，起初的设计与开发，以相关性、召回、准确为指导方针。系统实验之后，根据 CTR/CPM 以及转化的统计来对数据进行指导。最终的目标是提升整个内容页广告的质量和广告主的 ROI。

第三章 基于关键词聚簇技术广告检索需求

3.1 广告检索端需求分析

3.1.1 广告平台需求分析

广告检索平台，需要面对多种不同类型的用户。主要需要协调网站主、广告主和联盟的利益，最终形成一个生态圈。而检索端主要的作用是根据页面的特征，进行广告的检索并且返回展现。同时，检索端还有一系列的需求，包括平台需求、对广告主的业务需求等等。这里根据整个广告平台的用例，来分析检索端的需求。

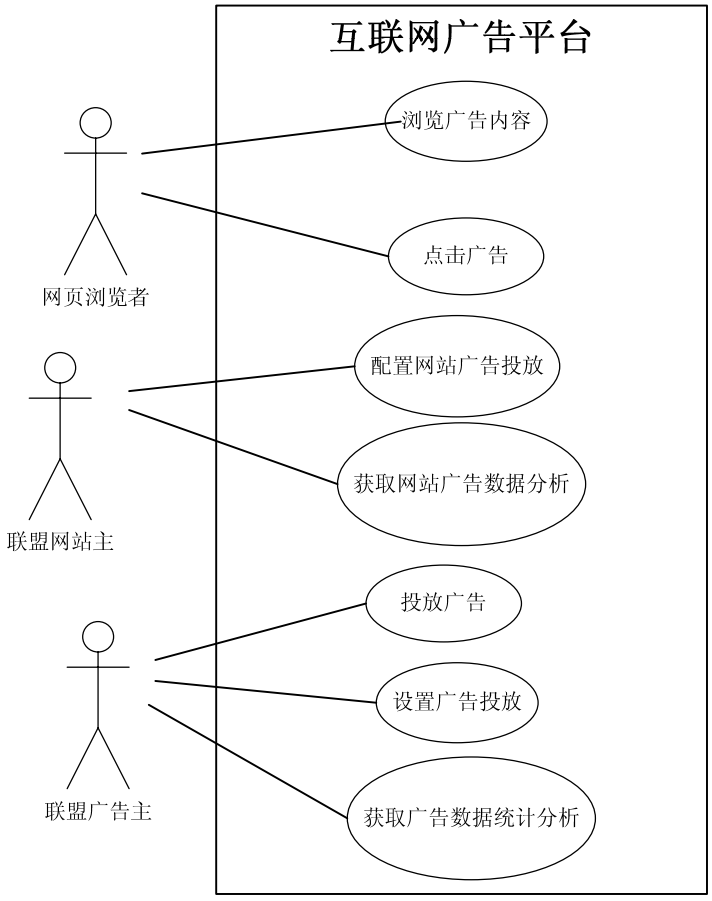

图 3.1 广告平台用例图

对整个广告平台来说，主要的参与者有网页的浏览者、网站主和广告主。同时当然还包括平台主，这里不进行标明了。分析各个参与者：

1. 网页浏览者：浏览网页，在浏览的同时，会有点击广告并且获取广告信息的需求。

2. 网站主：希望架设平台的广告资源，通过自己拥有的流量，进行变现。
3. 广告主：系统参与广告平台的广告投放，将自己的广告被更多的受众看到，期望为自己带来更多的利益和收入。

所有的参与者，通过前端的页面同检索端进行交互。这里对几个用例做说明，首先是与网页浏览者相关的用例：

1. 浏览广告内容：根据前端的请求（注意，如果浏览的是内容页则走这条分支），从广告库中进行广告检索。并且进行排序，返回广告。
2. 点击广告：系统需要对广告的点击动作做响应。

与网站主相关的用例：

1. 配置网站广告投放：如果网站主想要参与到平台的广告投放中来，可以通过该功能选择进行平台的广告投放。
2. 获取网站广告数据分析：系统能够对网站的所有展现、点击数据进行统计，方便网站主管理网站和获取变现数据。

与广告主相关用例：

1. 投放广告：广告主选择在平台上进行广告投放
2. 设置广告投放：广告主可以对广告的内容进行设置，包括广告的拍卖词、物料等等。
3. 获取广告数据统计分析：系统对所有广告信息进行统计，每个广告被点击、计价都会被系统记录并且反馈给广告主。

3.1.2 基于关键词聚簇的广告检索需求分析

基于关键词聚簇的广告检索方法，本质上是内容页广告检索的一个分支。不同的检索分支，共同构成了检索端的处理逻辑。这里通过一个流程图来描述该检索分支的逻辑：

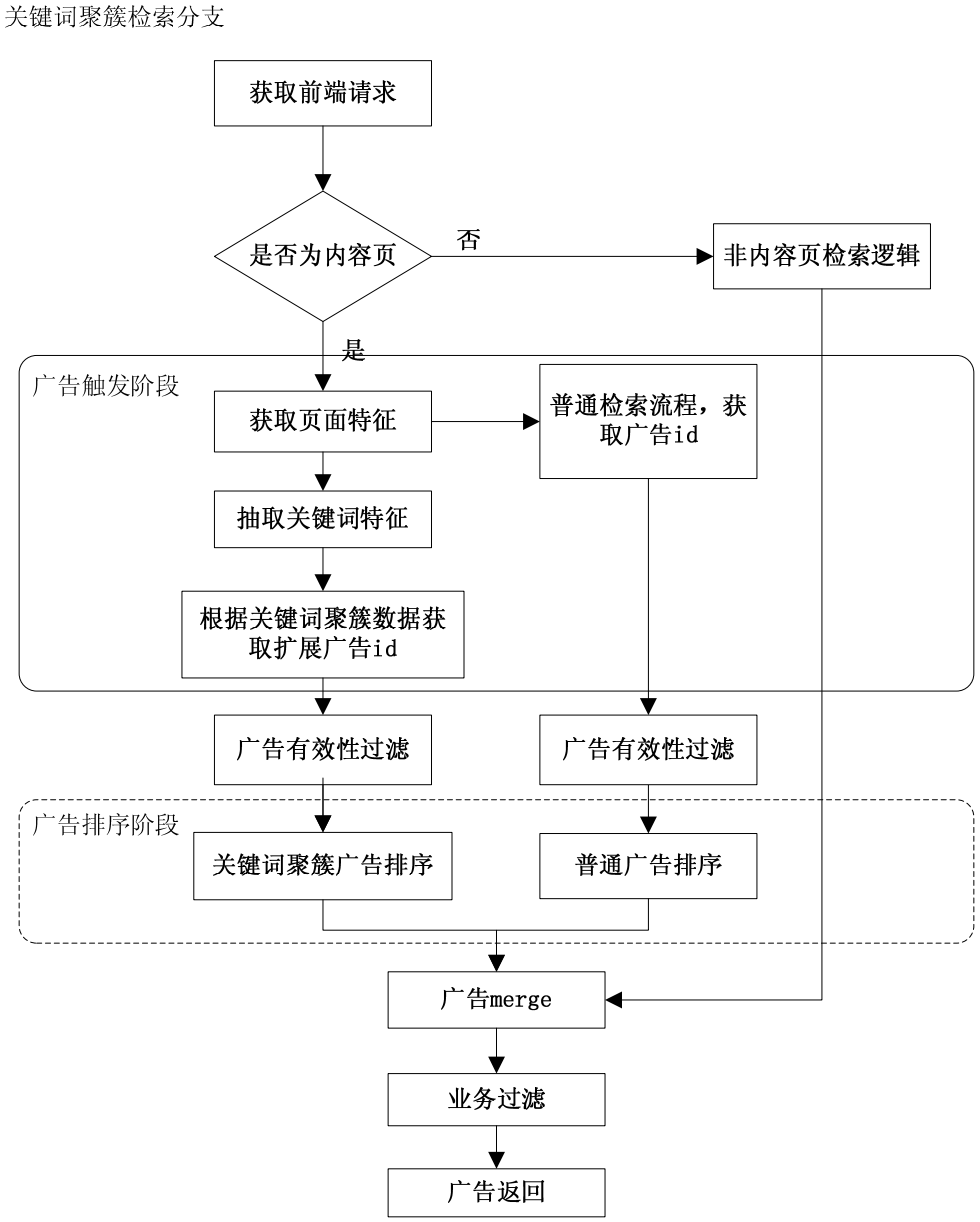

图 3.2 关键词聚簇检索分支流程图

关键词聚簇检索分支面向的对象是一次广告请求，广告请求所需要的，就是能够根据特征获取到相关的广告。从整个检索分支上看，主要有一下几个部分：

1. 相关性广告触发截断：根据特征，尽一切可能获取最大的广告返回集合。这个过程是指触发，后续会着重介绍触发。

2. 广告排序截断：能够根据相关性分数进行广告排序。后续会着重分析广告排序需求。
3. 广告 merge：将各种检索分支所触发的广告 merge 在一起，通过 ctr 预计以及广告价格排序。
3. 广告返回：将符合要求个数的广告返回给前端。

3.1.3 内容页广告触发要求

内容页广告的触发并不是一个非常复杂的过程，主要的几个要素是触发源、广告特征获取和业务过滤的需求，这里详细说明这几点的要求。

触发源：内容页广告检索，使用页面特征提取模块获取的页面触发 term 进行触发，触发词含有权重，根据不同触发词的权重对相应的广告进行不同权重赋值。

广告特征获取：每一个广告，在业务端的系统中，都含有广告的题词信息，这些信息用来表示广告的特征

业务过滤需求：在线上的系统中，广告特征不是静态的，是动态存在的。每一个广告都存在上线、下线的需求，也有可能余额不足，同时还会存在投放需求（例如仅仅投放在江苏地区等等），而对这些失效或者不满足需求的的广告，检索端会进行业务的过滤，最终得到的广告才会进行 GSP 计价并且最终展现。

3.1.4 触发排序的需求分析

排序一直是信息检索系统中比较复杂的部分，在广告检索系统中尤其如此。内容页广告的排序，分为广告的初选排序和优选排序。

初选也就是根据页面的触发 term，在广告库中获取广告（即召回）之后，对广告进行排序，这一阶段排序的主要作用是从召回的基础上，获取较为精准的广告，提高系统检索的准确率。

优选排序是跟在初选排序之后，对初选出来的所有广告，进行阈值的截断，对 top N 进行优选。这里主要考虑的原因是性能问题，优选往往需要考虑非常丰富的特征，比如展现/点击的预估、价格、历史表现等等，而初选的广告数量非常多，如果全部进行优选则性能难以承受。所以这里再对排序靠前的进行优选，

最终按照需求返回广告给前端，并且展现。

初选的依据是相关性，即计算一个广告和页面之间的相关性得分 p（ad |page）。在检索计算 p(ad|page)有很多方式，比如线性回归模型、分类算法等等。不管是线性还是使用一些分类的算法来进行计算，都需要对特征进行抽取，所以发现新特征、使用新特征是排序阶段比较重要的内容。

优选依据的内容就更加丰富，主要会使用到后验的很多数据来对模型进行训练，比如 ctr、cpm 统计预估等等。

在本文中，主要关心的是初选排序，因为加入了关键词聚簇之后，提供了新的特征，能够使得现有的排序模型更加完善。

3.2 自反馈体系

3.2.1 自反馈体系的介绍

在信息检索系统的发展中，一开始仅仅依靠简单的字面检索来进行内容的匹配。而后开始进入语义方面的检索，已经不简简单单进行字面的匹配，通过一些语义分析、语义匹配来指导信息检索。而在广告检索系统中，除了字面、语义，更重要的是一种意图的获取。互联网广告，依托于互联网的商务平台的成熟。一方面，成熟的商务平台可以提供交易、买卖、支付的功能，另外一方面，网络上巨大的流量，是一个非常巨大的展示平台。投放互联网广告，目的可以多种多样，但说到底，无非是为了“精准”投放，即最大限度的把广告推送给最合适的客户。但是目前的互联网技术，还无法完全真实的获得用户的意图。尤其在内容页广告上，如何解决这个问题是一个非常重要的突破点。

解决这个问题，最好的一个方法是让用户参与进来。即通过后验的挖掘，获取数据，再通过与用户的交互，来反馈给出评价。最终再反馈给检索端，通过点击、浏览数据统计的的方式来不断的完善数据字典。意图检索，就是通过“知识”来补充规则与推理。

举个例子，通过线下数据的训练，从大检索的摘要中可以分析出，“游戏”与“魔兽世界”（一款网络游戏）和“罗技 MX518”（一种游戏鼠标）都有一定的关系，从摘要中分析的相关度来看，“游戏”与“魔兽世界”的相关度要更加

的高。但是大检索的摘要，只能反映出在语义上两者存在的关联。通过意图学习，我们才能发现其中更深层次的关系。在一个游戏论坛的页面，页面的主题与“游戏”相关，于是，检索端通过“魔兽世界”和“罗技 MX518”分别进行广告触发，但发现，在不同的论坛页面，两者广告的点击率会有不同的表现，很多的页面上，“罗技 MX518”会比“魔兽世界”的点击率要高。于是，通过反馈，检索端会发现，在某些 URL 上，要相应的调整“魔兽世界”和“罗技 MX518”的权重，让后者的权重更高。这样能够获得更好的业务数据。

3.2.2 自反馈体系流程

检索端的自反馈目前主要还是针对展现和点击关系来进行。通过用户参与的点击行为，根据统计端的数据反馈，反馈出关键词簇的效果，对关键词聚簇的权值以及最终的排序数据进行校正。

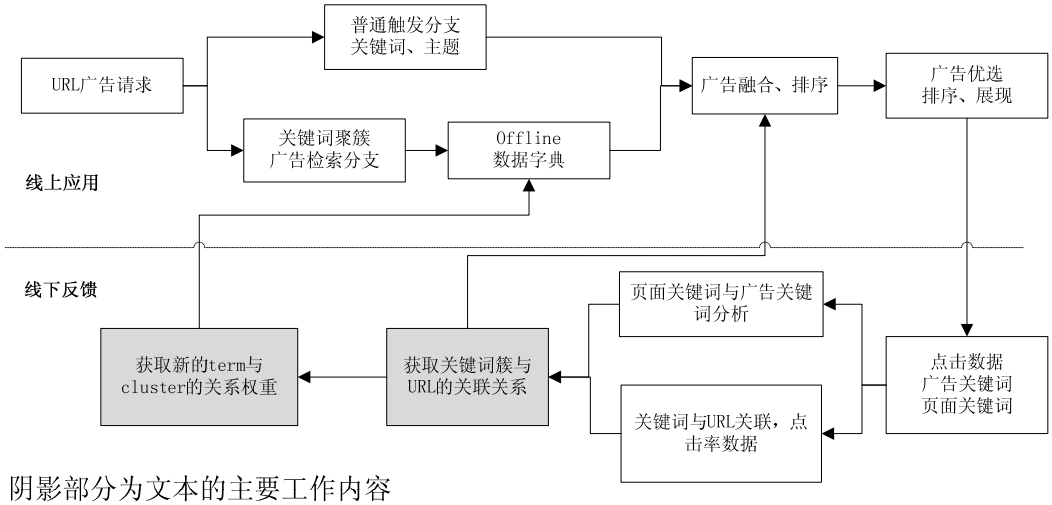

图 3.3 自反馈体系

通过点击数据的反馈，一方面，可以用 URL 与关键词簇的关联数据，来指导广告排序。另外一方面，可以通过这种后验的数据来指导关键词关系的挖掘。通过这样一套自我反馈的系统，让关键词聚簇的技术在内容页广告的检索端越来越准确越来越能符合用户的意图。

3.3 本章小结

检索技术的发展，都是从字面检索，到语意检索，再发展到意图检索。意图

检索最核心的思想就是用“知识”来补充规则与推理。检索端着重要解决的问题是，通过何种方法能够更好的去解决在检索端中触发特征单一稀疏的问题。关键词聚簇技术，可以很好的解决这个问题，他轻量级、自反馈，通过这种方法在检索端为内容页广告的相关性以及业务数据作出贡献。

第四章 检索端的总体设计

4.1 百度网盟检索端模型设计

4.1.1 整体架构

百度网盟的整体架构为一个星状结构。以广告检索模块为中心，分散各个功能模块。整个架构每天承载的流量在几十亿左右，需要使用一些分布式处理的架构来支撑，这些部分与本文关系大不，这里不进行详细介绍。

网盟检索端的流量是通过前端通过嵌入在 HTML 中的代码触发的。在网页加载的时候，发送请求给检索端，检索端通过一系列复杂的流量分析以及检索策略进行广告的选取[张友生，2009]，最终将广告返回给页面。检索端的整体检索流程以及通过的模块组合如下：

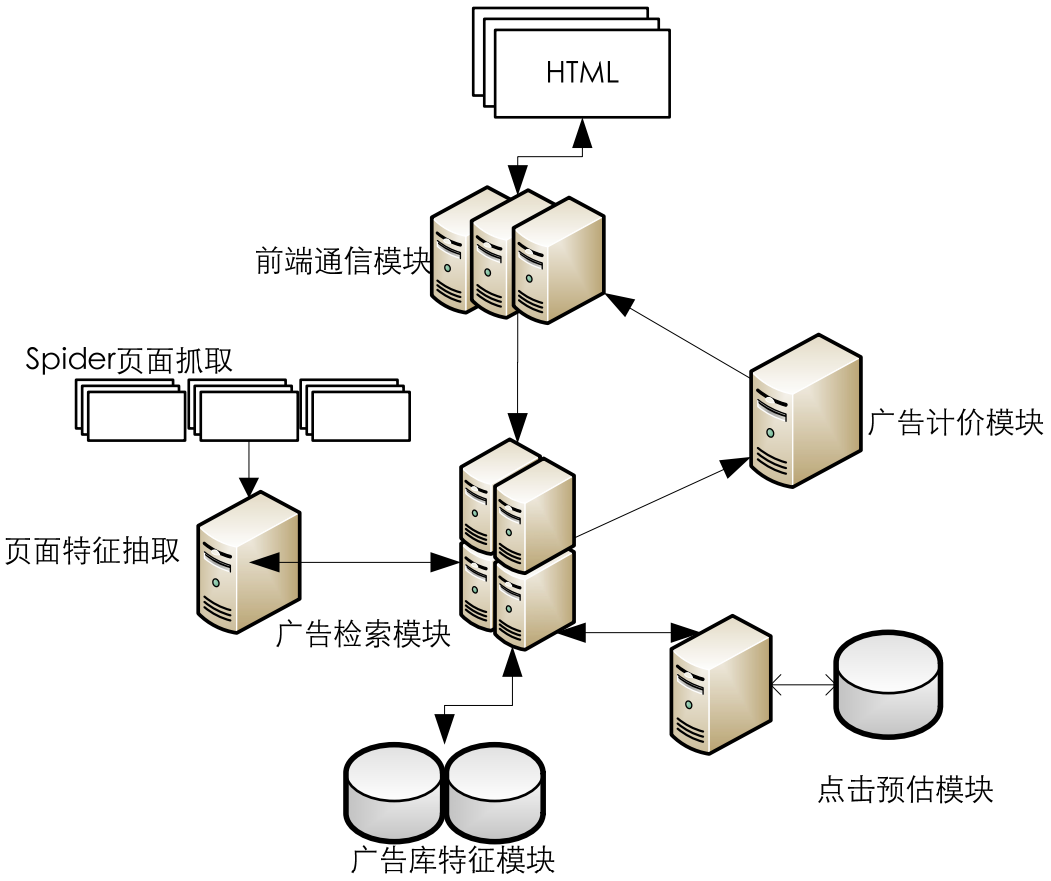

图 4.1 检索端整体架构

4.1.2 模块简介

检索端主要的模块如上图，分别有：前端通讯服务模块、广告计价模块、页面特征提取模块、广告检索模块、点击预估模块、广告库特征模块。在这里，具体介绍下各个模块的主要功能：

前端通信模块：通过嵌入在 HTML 中的代码，在页面浏览的同时会像检索端发送请求，通过通信模块，对前端的请求进行封装和处理，组成可以被检索端处理的请求格式，发送给广告检索模块。在经过计价处理之后，再交由该模块返回给前端页面展现[Bruce，2009]。

广告计价模块：计价直接关系到广告平台的收入，需要使用单独的计价模块来对广告进行点击计价并且对一些作弊流量进行过滤。

页面特征提取模块：系统会通过爬虫对互联网页面进行抓取，将网页存库。在该模块中，对网页特征进行抽取、精选、封装，最终将特征形成可使用的关键词特征、主题特征、分类特征。

点击预估模块：该模块连接在广告检索模块中，能够获取到广告检索中所有可使用到的特征，包括前端的请求特征、用户的特征、广告库特征、页面特征等等，综合这些特征对广告的点击率做出预估，在高级排序这个环节中，点击预估的结果是一个重要的标准。

广告库特征模块：广告库维护了当前系统所有可以使用的广告，在维护广告的可用性的同时，还对广告的特征进行抽取，供检索端使用。

广告检索模块：是整个检索端的核心模块，获取到页面特征提取模块获取的特征，连接广告库，根据广告特征与迎面特征的匹配程度，再根据不同广告请求的业务特点，抽取广告。然后进行广告的相关度排序，最后相关性排序最高的几条广告返回给计价模块。

4.2 该技术应用的模块

4.2.1 关键词聚簇应用于触发

通过线下挖掘获取到关键词簇字典，将在检索中会使用到的簇之间的关系描述分为两部分：

term→cluster：从一个关键词出发，获取该关键词所属的关键词簇，并且通过赋权计算，获取一个 term 与该簇之间的权重关系。

cluster→cluster’：从一个关键词簇触发，获取与之相关的关键词簇，通过对两个关键词簇的支撑词进行匹配计算，获取一个关键词簇与扩展关键词簇之间的权重关系

上述两个索引，维系了从一个触发词触发，所能获取到所有的关键词簇。另外，需要在广告库中维护广告与簇之前的关系。每一个广告，都含有广告的拍卖词，对广告的拍卖词进行切词处理，再计算权重，最终获取了每一个广告与簇之间的关系，以一个广告索引的方式存在于广告库中，在系统中，称为关键词簇索引：cluster→ad

在广告检索模块中，将 term→cluster 与 cluster→cluster’两条索引加载到内存。当前端传来 URL 请求，依次经过服务器与网页特征提取模块之后，获取到页面的特征关键词（term），接着根据 term→cluster 索引获取初始 cluster 集合，再根据 cluster→cluster’索引获取扩展 cluster 结合，最终，根据所有的 cluster 集合，在 cluster→ad 的倒排索引中检索广告，形成初步的触发广告集合。

4.2.2 关键词聚簇应用于排序

广告排序最终需要得到一个广告同一个页面之间的关系，这里用一个概率模型 p(ad|page)来描述，即一个广告和这个页面相关的概率[许家梁，2009] 。在检索端中，将页面抽象成许多的特征，每个特征都部分的代表了页面。关键词聚簇技术，主要考虑的是页面的关键词特征。一个广告同所有的特征关键词的关系，就描述了该广告同页面之间的关系。如何计算这个一条广告同所有特征关键词的关系，可以简单的描述为：单特征关键词下广告权重取最大，多关键词下广告权

重求和。基于关键词聚簇检索的过程是一个比较简单的图遍历，可以分几步来描述下：

第一步：通过页面特征分析模块，获取所有的特征关键词。

第二步：根据关键词簇在广告集合中获取广告。从页面特征模块可以获取到页面的特征关键词，每个词都含有权重 w(term)。从该关键词触发，利用之前描述的三条索引，从 term→cluster 出发，获取所有的初始 cluster 集合。根据初始 cluster 集合，遍历 cluster→cluster'获取扩展 cluster'集合，将初始 cluster 同 cluster'组合在一起成为关键词簇集合。

第三步：根据 cluster 集合，遍历 cluster→ad 倒排。则得到最终广告候选集，其中，一条广告被一个触发词触发的关系用一条路径权重来描述：

p(ad|term_1) = w(term_1)*p(cluster|term_1)*p(cluster'|cluster)*p(ad|cluster')

第四步：在一个 term 下，一个广告可能被多条路径遍历，这里反映的是一个广告同一个 term 的关系，所以在不同的路径下，p(ad|term)被描述为多条路径中权重最高的一条：

p(ad|term_1) = w(term_1)*MAX(p(cluster|term_1)*p(cluster'|cluster)*
p(ad|cluster'))

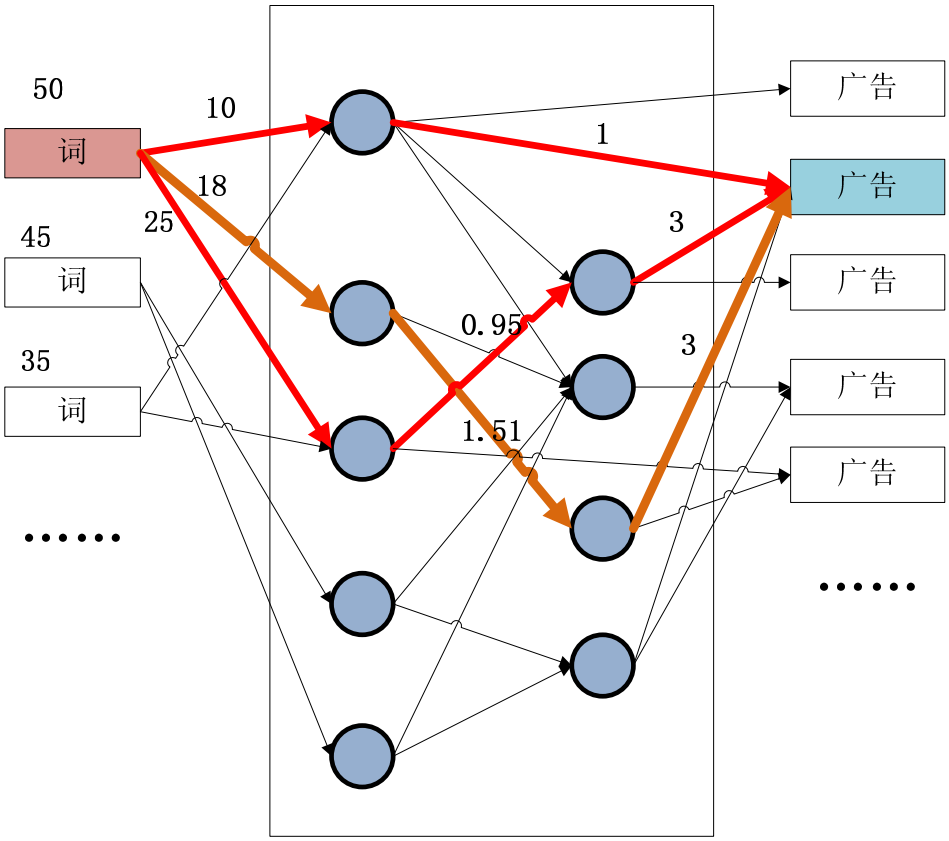

图 4.2 单关键词广告权重计算

如上图中，第一个关键词的权重为 45，从他出发，有三条路径可以到达同一个广告，取这三条路径中权重最大的一条路径，为该关键词和这条广告的权重。即图中橘黄色的那条路径，p(ad|term)=50*18*1.51*3。

第五步：在关键词聚簇广告的检索中，可能会出现不同的 term 都有路径指向一个广告。这里反映的是一个页面关键词集合的主题集中性，反映了一个广告同页面之间，关键词聚簇的效果，这里在计算权重的时候，考虑多路径之间的求和：$p(ad|page) = \sum_{k=0}^{n} p(ad|term_k)$。则这里最重获得权重 p(ad|page)为广告排序最终获得的权重。

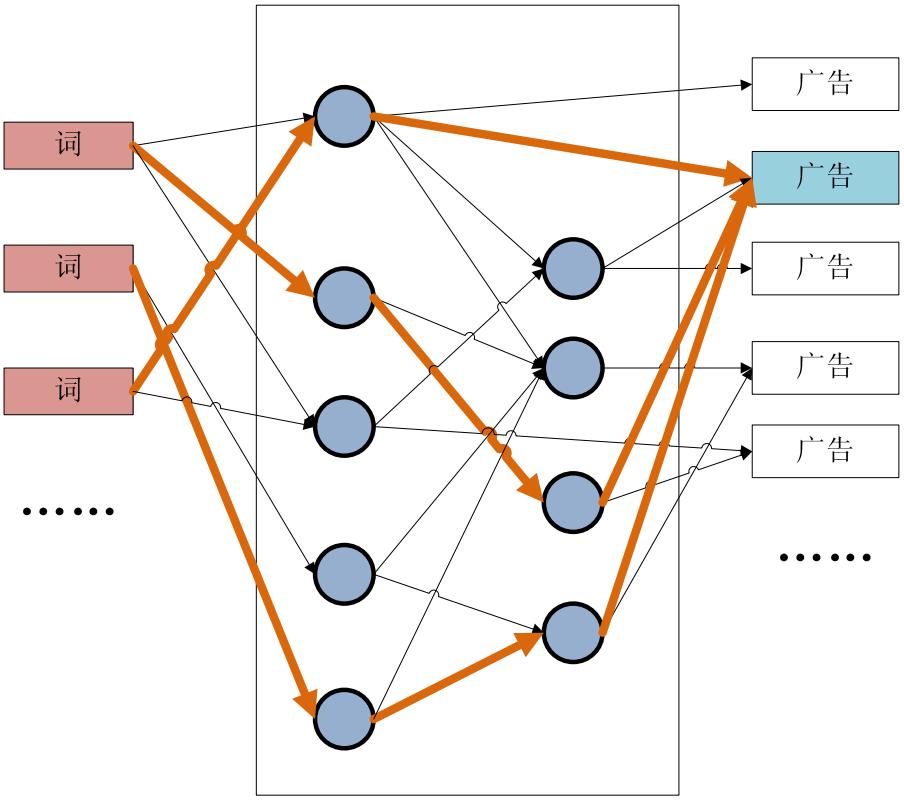

图 4.3 多关键词命中同广告权重计算

如上图中，对于同一条广告，三个触发词都能够找到一条最佳路径达到同一条广告，对于这条广告最终的权重，就是这三条路径权重之和。

4.2.3 如何通过参数体系调整触发和排序

检索端是一个参数体系，需要能够通过对不同参数的调整，灵活的改变检索端广告检索的状态。使得检索端在广告的召回、准确、想关度之间灵活的变动，以满足不同情况下不同的需求。关键词聚簇技术同样在这个参数体系的控制之下，通过这个参数体系，由关键词聚簇所检索出来的广告，可以灵活的在召回、准确之间进行调整。同时还可以通过系数来保证整个排序的效果。目前在系统中，使用到的与触发排序相关的参数以及这些参数的描述：

term_cluster_thd：在进行 term→cluster 索引遍历的时候，对 term 和 cluster 的权重进行截断。

cluster_cluster_thd：在进行扩展遍历的时候，对扩展的 cluster 进行权重的截断。该参数和之前的 term_cluster_thd 两者对关键词聚簇的数据进行约束，能够使得该体系的调整非常灵活，如果需要高相关性的广告，那么将这两个系数调高。

如果广告召回少，那么将这两个系数降低，这样能获得较多的广告召回。

cluster_score_param：在相关性计算体系中的系数（相关性计算体系介绍见4.4.2）

cluster_unit_num_thd：最终使用关键词聚簇对广告触发后，对广告的个数进行截断。根据相关性分数排序，取权重最高的 N 个广告返回。

4.3 对触发以及排序的影响

4.3.1 相关性计算体系

相关性计算，本质上就是要根据页面的特征和广告的特征，获取一个广告和页面之前的相关性评分：p(ad|page) 。目前的相关性计算体系，是一个多特征的线性回归模型，目前使用的特征有关键词特征、主题特征和分类特征[王知津等，2005]。

关键词、主题和分类都是离散化的数据，大体的格式都是“内容 A、权重 B”，在进行相关性计算的时候，从页面特征处，可以获得页面的计算特征序列，包括页面的关键词、主题和分类，类似上述的格式。同时，在触发的时候，还能获得广告库中广告的计算特征序列，包括广告的关键词、主题和分类。当获得了两端的特征序列的向量之后，使用内积的方式分别计算各个特征的权重。

例如：

从页面获取的关键词特征是：A|25，B|15，C|5

广告的关键词特征是：A|19，D|8

则计算该广告在该页面展现的分数是，由于仅仅有关键词 A 两者是匹配的，所以在关键词这个维度上，计算的分数为：25*19=475。使用同样的方法，分别计算主题特征和分类特征维度上的分数，最终获得了广告在三个维度上不同的表现分数。

有了初始的分数计算体系，我们需要目标函数。通过 PM 的标注样本集，然后将这三个分数，使用线性拟合的方法，确定各个维度的参数，并且得出最终的相关性分数[杨小平等，2003]：

$$p(ad \mid page) = \alpha * term_{score} + \beta * topic_{score} + \gamma * category_{score} + N$$

4.3.2 融入相关性计算模型

原先的相关性计算，通过 term、主题、分类这三维特征计算向量内积，再通过对样本训练获取的结果，使用线性回归模型，获取最佳的系数。现在通过获取了 cluster score 之后，将该分数也加入到原先的分数特征中，将这四维组合在一起，再通过线性回归模型获取最佳的参数组合。则现在的相关性分数是：

$$p(ad \mid page) = \alpha * term_{score} + \beta * topic_{score} + \gamma * category_{score} + \delta * cluster_{score} + N$$

广告在计算了相关性分数之后，再对分数进行截断。关键词聚簇模型融入相关性模型之后，能够使用之前的一套训练体系来进行参数训练，同时使用统一的截断来控制广告召回的数量和广告相关性的准确性。

在系统中，关键词、主题、分类这三维都存在一定的缺陷：

关键词特征：关键词特征直接来自于页面，从页面中提取关键词，经过赋权得到，但是从页面获得的关键词往往比较稀疏，并且还有一些比较明显的噪音。比如页面广告、网络留言等等。

主题特征：与主题模型相关，目前的模型还不能取得非常好的效果。

分类特征：分类特征过于粗糙，分类的粒度较大，这样在使用的时候，很没有区分度。

而关键词聚簇这一维度，通过关键词的主题集中性来解决了关键词特征中的噪音问题，同时对关键词的主题有一定的修正，很好的弥补了之前三维特征有所缺失的信息。使得相关性体系更加的完备。

4.4 本章小节

本文主要介绍的是网盟中相关性广告检索架构，基于关键词聚簇手段的检索，如何融入这样架构的方法。在广告触发环节中，通过加载自动化生成的索引结构，基于页面的关键词特征进行，通过广告库中的倒排索引进行广告的触发。在排序中，关键词聚簇将重点放在词的内聚性上，使用单路径求最大，多路径求和的方法。挖掘出一个广告和页面关键词之前的关键词内聚的权重，并且基于该权重为广告赋权。并且基于目前存在的相关性模型，将该特征融入其中，获得一个比较好的效果。

第五章 模型应用与实现

5.1 检索端的实现

5.1.1 模型建库方案

在设计之中，使用了三条拉链，分别是：term→cluster、cluster→cluster 和 cluster→unit，在检索端，需要将这三条拉链载入内存，提供倒排给检索广告使用。在 load 的过程中，需要综合考虑在检索端的平均响应时间和内存以及 CPU 资源的消耗上。三条索引都是倒排的结构，即从一个特征触发，找到响应的节点，并且每个节点都有权重含义，可以定义为<A,[B|weight]>的形式[李建中 等，2006]。

这三个索引，cluster 和 ad 都是以 id 的形式出现，使用三十二位的无符号整型保存，并且有一个权重也是三十二位的无符号整型。节点的数据类型如下：

```
typedef struct _cluster_node_t {
    // 目前有三套索引
    // term->cluster:
    //     index_value : cluster id
    //     index_weight : cluster weight
    // cluster->cluster:
    //     index_value : cluster id
    //     index_weight : cluster weight
    // cluster->unit :
    //     index_value : unit id
    //     index_weight : unit weight
    u_int index_value;
    u_int index_weight;
} cluster_node_t;
```

图 5.1 簇节点结构

上表是节点类型的定义。为了使得索引的遍历速度加快，cpu 的 cache 命中非常重要，倒排的结构需要使用连续的数据结构，在内存中，维护一个 cluster_node_t 的数组。然后通过 hash 的形式<key,value>[李建中 等，2006]，使用一个 key（可能是 term、cluster），找到相对应的一组连续节点的起始位置和结束位置，进行遍历，结构如下图：

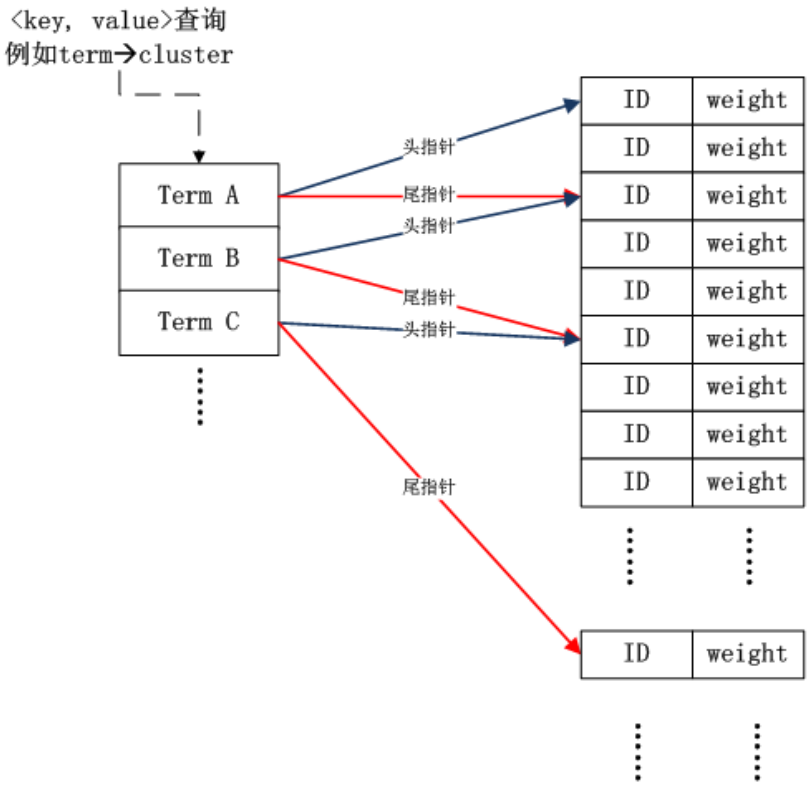

图 5.2 倒排索引的遍历方法

这样遍历索引的时候，使用如下的遍历逻辑：

使用某条索引的引用（term→cluster、cluster→cluster'、cluster→unit）

使用该索引的 key，遍历索引的 hashmap，根据 hashmap 在 node_array 中找到拉链，伪码如下：

```
IF key in hashmap THEN
    从 hashmap 中获取头节点 idx_head，尾节点 idx_tail;
ELSE key not in hashmap THEN
Return;

For i=idx_head, i<idx_tail, i++
    Get node_array[i];
```

图 5.3 索引遍历伪码

term 比较特殊，在检索端使用对字符串签名的方式来保存 term，格式是两个无符号整型。Hashmap 使用了一个专门为这种签名形式设计的 dict_search_hash_t，他的描述为<<sign1,sign2>,<cuint1,cuint2>>：

```
typedef struct dict_search_hash_head_t
{
   unsigned int sign1; //签名1
   unsigned int sign2; //签名2
   unsigned int cunit1; //索引中的头结点
   unsigned int cunit2; //索引中的尾节点
} dict_search_hash_head
```

图 5.4 字典查询数据 dict_search_hash_t

这样，最终的整个索引 cluster_index_group_t 定义完毕，如下：

```
   typedef struct _cluster_index_t {
      u_int node_num;   //节点数据的个数
      u_int node_size;  //系统所能容乃节点个数的上限

      cluster_node_t *cluster_array;
      // sign1: term sign1, sign2:
      //       term sign2, cuint1:
      //       head index, cuint2:
      //       tail index
      // term->cluster :
      //     sign1 : term sign1
      //     sign2 : term sign2
      // cluster->cluster :
      //     sign1 : cluster id
      //     sign2 : 0
      // cluster->unit :
      //     sign1 : unit id
      //     sign2 : 0
      // search dict, just use to search
      dict_search_hash_t *index_search_dict;
   } cluster_index_t;

   typedef struct _cluster_index_group_t {
      cluster_index_t *term_cluster_index;    // term->cluster
      cluster_index_t *cluster_cluster_index; // cluster->cluster
      cluster_index_t *cluster_unit_index;    // cluster->unit

      time_t last_upd_time;  // 用于reload时候的时间判断
   } cluster_index_group_t;
```

图 5.5 索引集合定义

5.1.2 触发排序的实现

之前已经介绍了触发排序的设计以及实现：

触发阶段：主要的触发逻辑是根据触发词，遍历三条关键词聚簇线下模块生成的索引（term→cluster/cluster→cluster'/cluster→ad），获取广告，并且在遍历广告的同时，就能得出相关性广告的得分（cluster score），并且将它最为广告的特征传递到排序阶段。

以下是部分代码，主要的功能是在触发之前做的准备工作：

```
#include "mod_trigger_cluster.h"

INIT_MODULE_DECLARE(trigger_cluster)
{
   ……
   /// 双buff加载关键词聚簇数据
   load_cluster_group(index_buff[2],
                                    cluster_index_path);
   ……
}

RELOAD_MODULE_DECLARE(trigger_cluster)
{
   if (ret < 0 || !(st.st_mode & S_IFREG)) {
       BS_LOG_WARNING("fail to stat path[%s]", fname);
       return RUN_FAIL;
   }

   ……
   if (st.st_mtime > g_cluster_index_group->last_upd_time) {
       /// 重载索引数据
       ret =
       NOVA_BS::load_cluster_group(
              &g_trigger_cluster_conf.index_buff[buff],
              g_trigger_cluster_conf.cluster_index_path);
   }

   ……
   return RUN_SUCC;
}
```

图 5.6 触发前数据准备

以下代码主要是检索执行前和检索时候的部分代码：

```
BEFORE_SEARCH_DECLARE(trigger_cluster)
{
   /// 获取控制数据
   as2bs_ctrl_t *p_as2bs_ctrl = NULL;
   if (c_d->get_asreq_ctrl(p_as2bs_ctrl) != RUN_SUCC) {
       BS_LOG_FATAL("c_d->get_asreq_ctrl failed.");
       return RUN_FAIL;
   }

   /// 获取关键词特征
   try {
       as2bs_req_t *p_as2bs_req = NULL;
       if (c_d->get_asreq_idl(p_as2bs_req) != RUN_SUCC) {
           BS_LOG_FATAL("get_asreq_idl fail");
           return RUN_FAIL;
       }
       const fea_header_t cr_fea_header = p_as2bs_req->fea_header();
       ……
   }
   catch(const bsl::Exception & cre) {
       return RUN_FAIL;
   }
   catch(...) {
       BS_LOG_FATAL("%s", "sth bad happened");
       return RUN_FAIL;
   }
   return RUN_SUCC;
}

EXEC_SEARCH_DECLARE(trigger_cluster, s_d, c_d, m_d)
{
   ……
   /// 关键词聚簇触发调用，并且将广告触发加入广告队列
   if (trigger->kc_trigger_process() != RUN_SUCC) {
       BS_LOG_FATAL("trigger run_trigger_process fail!");
       return RUN_FAIL;
    }

   return RUN_SUCC;
}
```

图 5.7 检索流程代码

排序阶段：根据广告的特征集合，计算出广告各个维度的得分，并且得出最终的相关性评分。以下是整个触发以及排序的流程，标星的部分联系了触发和排序阶段。

图 5.8 关键词聚簇技术触发与排序流程描述

5.2 参数体系的实现

在这样大型的广告系统检索端，参数系统的实现是非常重要的。整个系统的架构从模块上看含有多个模块，一次请求将会经过不同的模块，最终返回到前端。通过对一次请求的流量进行标示，可以获取到该流量的状态。而一次请求的状态里面，最重要的是与该请求相关的参数集。

在多个模块配合完成功能的检索端中，参数体系非常重要的一个需求是能够“各取所需，互不影响”。因为参数的传递是关系到接口的，如果某模块需要从上游模块获取某参数，那么他必须要修改模块本身（修改接口以及添加参数处理逻辑），同时，他还需要修改上游模块的接口，使得上游能够获取这个参数，并且将它传递给你。这样，为了修改自身的参数，需要对别的模块造成影响。耦合很大，不是一个很好的设计。

在检索端中，使用的网络数据流传递参数的形式。将所有的参数打包成为一个字符串流，通过 json 的形式[百科，2011]，在各个模块中传递。JSON(JavaScript

Object Notation) 是一种轻量级的数据交换格式。易于人阅读和编写，同时也易于机器解析和生成。它基于 JavaScript（Standard ECMA-262 3rd Edition - December 1999）的一个子集。 JSON 采用完全独立于语言的文本格式，但是也使用了类似于 C 语言家族的习惯（包括 C, C++, C#, Java, JavaScript, Perl, Python 等）。这些特性使 JSON 成为理想的数据交换语言[IBM，2008]。如果某模块需要该参数，则直接从 json 流中获取参数。

声明所有的参数集合：

```
typedef struct _cluster_param_t
{
    u_int term_cluster_thd;
    u_int cluster_cluster_thd;
    u_int cluster_score_param;
    u_int cluster_unit_num_thd;
    float cluster_score_param;
}cluster_param_t;
```

图 5.9 关键簇实验参数说明

通过在配置文件中进行 key/valued 的配置，通过 json 的封装，将参数封装成 json 的格式：

```
# cluster config
term_cluster_thd :    20
cluster_cluster_thd :    30
cluster_score_param :    0.002
cluster_unit_num_thd :    140
cluster_score_param :    0.05
```

图 5.10 参数配置方式

当下游的模块获取该字符流之后，通过以下函数调用，就可以获取到参数中的内容，这样省去了模块之间接口的作用：

```
cluster_param_t* cluster_param = json.parseByString(config_string);
```

图 5.11 参数获取方式

5.3 性能设计

5.3.1 性能问题

检索端每天需要面对几十亿的流量，对检索端模块的性能要求非常高。从几个方面上综合考虑各项性能指标。

平均响应时间：对每条请求，检索端处理一次所需要花的时间。需要在毫秒级别处理完毕。

内存使用情况：模块的机器所使用的内存在 40G 左右，添加了新的逻辑，需要保证内存的总使用量不会达到危险线（大约为机器内存的 90%，即 40*0.9，36G 以下）

CPU idle：大量的逻辑运算会导致 CPU 全程中高速运行，CPU idle 表明了 CPU 剩余可以用的空间，如果 idle 长期保持为 0 的话，那么机器很有可能会宕机。

平均每秒可承受的响应数量（request per second）：模块使用多线程处理请求，线程的同步问题会导致每秒的请求受限，该指标是一个辅助指标。

表 5.1 初始的性能数据

性能指标	平均响应时间	内存使用情况	CPU idle	request per second
原性能	4ms~5ms	28G	35%	900~1000
先性能	6ms~7ms	29G	15%	800~850
性能变化	+50%	+3.5%	-57%	-11%

从这份数据上看，关键词聚簇技术加入到检索端之后，平均响应时间和 CPU idle 有较大的下降。而由于线程的平均响应时间增长，导致每秒所能能处理的请求个数也有较大幅度下降。

主要分析的原因在两点：

1. 三条索引存在冗余，比如 term→cluster 中不存在的 cluster id 还被 cluster→cluster'索引等等，冗余数据导致遍历、去重的时间提高给多，CPU 需要消耗更多的计算。
2. 在检索广告的时候，会考虑三条索引，每条索引都会使用 hash 结构来保存索引位置，导致在检索的时候 hash 表的检索时间非常长[李师贤等，2003]。举个例子，如果一个页面有 30 个触发关键词，平均每个 term 会根据 term→cluster 找到 100 个 cluster，每个 cluster 会根据 cluster→cluster' 找到 50 条 cluster'，最后所有的 cluster 需要遍历 cluster→ad，则这个过程中，总共需要查 hashmap 的次数为：30*100*50，在十万的量级，一次检索就需要十万次的 hashmap 的查表字数，对性能的影响较大。

5.3.2 模型与数据结构的优化

性能问题需要解决上述所说的两个问题：

1. 数据冗余的问题：通过对 term→cluster/cluster→cluster'/cluster→ad 进行索引，将 term→cluster 中不含有的 cluster id 从第二条索引中去掉，将不会被遍历到 cluster id 从 cluster→ad 中去掉，这样数据量会下降很多。最终的结果显示大概能够将数据文件的大小缩小一半。
2. Hashmap 查找次数过多的问题，由于 hashmap 中存在命中率的问题。所以处理这个问题的方法有两点：
   a) 重新计算 hashmap 中桶的个数，根据数据规模的大小调整桶的数量，提高 hashmap 查找的命中率
   b) 将三条索引合并为两条索引，term→cluster/cluster→cluster'本质上是要获得从 term 到 cluster id 之间的关系，只是无非是扩展和非扩展的区别。通过线下的计算，可以把扩展和非扩展一视同仁，合并为一条索引。即将 term→cluster/cluster→cluster' 合并为一条索引 term→cluster''，其中 cluster''包含扩展和非扩展的所有关键词簇。这样在遍历的时候，则省去了 cluster→cluster'这个索引 hashmap 的查表操作。

经过这两个步骤的优化，性能问题得到缓解，优化之后的性能数据如下：

表 5.2 优化之后的性能数据

性能指标	平均响应时间	内存使用情况	CPU idle	request per second
原性能	4ms~5ms	28G	35%	900~1000
Old 性能	6ms~7ms	29G	15%	800~850
New 新能	5ms	28.7G	31%	900~100

5.4 业务指标的提升

5.4.1 最终业务指标

前文提到，最终的业务指标需要从多个方面来评价关键词聚簇模型对检索系统带来的升级。如何评定最终的效果，需要根据线上小流量实验的对比数据分析获得。主要考量的指标有：

1. 广告的转化率，这里通过业务端的到达率（到达的概念是产生点击之后，广告的页面被用户浏览。到达率是指到达的次数同点击次数的比率。这是衡量展现广告质量的一个指标）和二跳率（二跳率的概念是当网站页面展开后，用户在页面上产生的首次点击被称为“二跳”，二跳的次数即为“二跳量”。二跳量与浏览量的比值称为页面的二跳率。这是一个衡量外部流量质量的重要指标）数据来评定[朱胜 ，2009]。
2. 业务数据上的提升，包括 CTR/CPM。
3. 广告的相关性提升。

最终进行流量对比实验之后，从三个指标上看都有提升。从三个方面分开分析：

1. 广告转化率：到达率+3.1%，二条率+4.2%
2. 业务数据：CTR1+1.5%，CPM1+5.3%
3. 广告相关性结论：随机抽取 200 个 URL，其中关键词聚簇影响的 URL 在 60 个左右，PM 评估样本，得出的结论是 A+B 类 case 提升 21%，D+E 类 case 下降 29%，对整体的相关性提升较明显：

表 5.3 相关性评估结论

1 档	2 档	3 档	4 档	5 档
1	10	23	4	1
2.56%	25.64%	58.97%	10.26%	2.56%
1	8	22	6	1
2.63%	21.05%	57.89%	15.79%	2.63%

5.4.2 业务指标的分析

从业务指标上看，引入了关键词聚簇的手段之后，从三个方面来分析，都给系统带来了较明显的提升。从根本上看，还是提升了相关性广告的占比。在内容页流量上，当相关性广告的比率提升之后，无疑对整体的点击率和转化率有提升。毕竟站点投放广告没有考虑到相关性。

从下图可以分析，关键词聚簇技术的引入，使得站点投放广告的占比下降较多。提升了整体流量上相关性广告的占比。

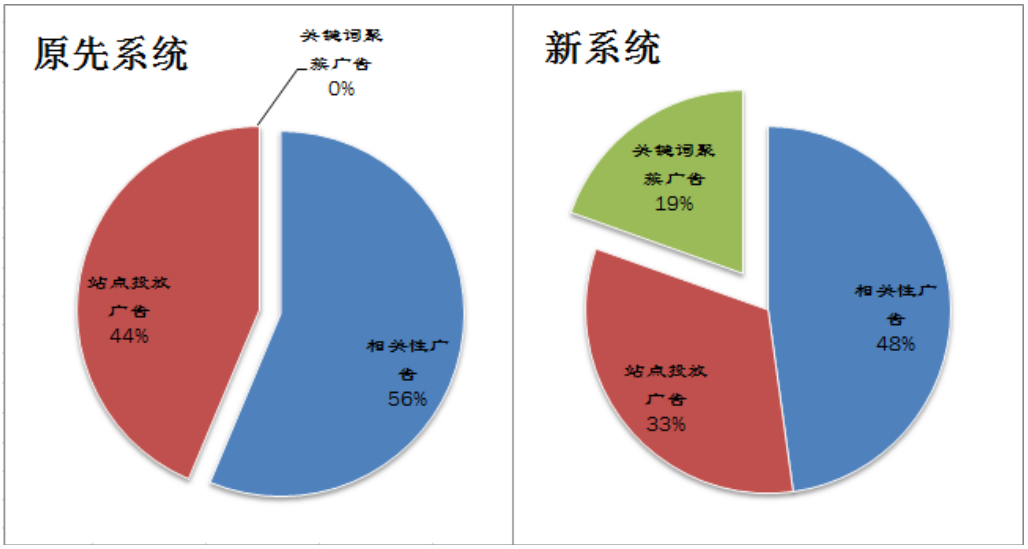

图 5.12 广告占比分析

另外单独评估关键词聚簇广告，同之前的相关性广告相比，相关性相差不大。这样关键词聚簇的广告的加入，让相关性广告得到更多的展现机会，摆脱了原先那种仅仅通过关键词匹配方式来触发广告的局面。

最后我们还同目前比较流行的一些主题类模型进行比较，经过比较，我们能够发现，基于关键词聚簇技术来进行广告展现，有一定的优势：

1. 召回率上考虑，主题类只能通过权重进行控制，但是关键词聚簇模型能够通过赋权和扩展两种方法进行控制。同时召回率上，关键词聚簇模型更佳，在归一化权重中，想通的截断阈值上，关键词聚簇模型的召回率要高于 plsa。

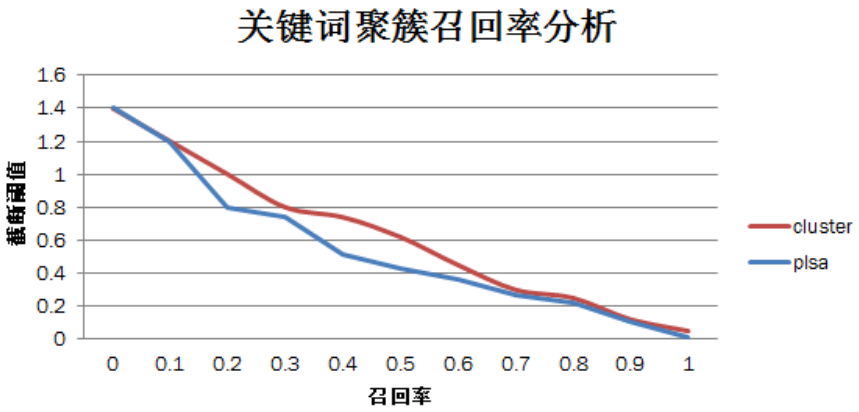

图 5.13 召回率分析

2. 准确率上考虑，主题模型更加通用，在内容页广告上，关键词聚簇模型能够达到更好的要求。所有的关键词聚簇数据都是与检索以及广告相关，能够更好的反映出内容也广告的特质。在准确率上，当截断高的时候，plsa 的相关性会非常好。但是，当考虑到召回而放开截断之后，plsa 的效果要比关键词聚簇技术差一些。

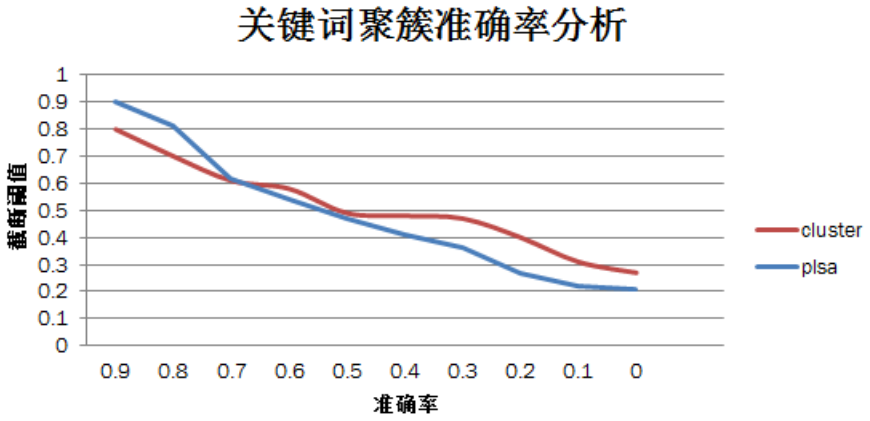

图 5.14 准确率分析

5.5 本章小节

内容页广告中，相关性广告和站点投放广告的比率很关键。但是相关性广告的资源并没有那么多，仅仅依靠关键字匹配，相关性广告的召回明显不够。在主

题模型受限的情况下，关键词聚簇应运而生，通过比较简单的线下模型的挖掘和训练，解决了一系列包括性能、参数控制问题，最终为广告检索端带来了新鲜血液，相关性广告的占比有较大幅度的提升，最终为系统的可持续发展打下了坚实的基础。

第六章 总结与展望

6.1 项目总结

关键词聚簇技术，从技术实现上，比较轻量级，通过线下挖掘，对关键词进行聚簇。通过关键词赋权，以及二元关系挖掘，给簇关系进行赋权。然后通过建库，引入内容页广告的检索端。

从历史数据分析，仅仅通过关键词部分匹配召回是远远不够的。我们需要在检索端引入语义和意图检索。通过 topic 模型，我们可以获取这样的效果。但是 topic 模型，由于算法的复杂度、模型训练的难度、通用性，导致在内容广告体系中，很难找到自己的位置，不能够很好的应用在内容页广告和用户意图检索中。轻量级的模型训练、自反馈是关键词聚簇的特点。通过利用一切可利用的资源，包括大搜索、广告库、业务系统等等，单机完成数据的预处理和挖掘，能够很快在系统中产生作用。同时，快速的利用反馈，让用户参与到模型的优化中来，通过点击数据和关键词反馈数据，进一步优化模型本身。为相关性广告引入新鲜血液，并且最终从业务数据、转化以及相关性分析上取得比较好的效果。

6.2 关键词聚簇技术的展望

关键词技术的第一次在内容页广告上应用，取得了比较好的效果。但是从理论模型上看，关键词聚簇技术还存在很多需要完善的地方。在内容页广告中，通过不断的反馈，关键词聚簇会做的越来越准，越来越专。与最终的浏览者意图越来越接近，通过这样的良性循环，为检索端带来更多的收益。

另外一方面，关键词聚簇模型虽然最初是设计在内容也广告中使用，但是它同样能够为搜索广告带来一定的收益，因为他们本质都是基于 Query 的关键词聚簇技术，该模型进一步挖掘了词和词之前的关系。除了搜索广告之外，由于它的轻量、可控，可以推广到多种环境中利用它来产生收益，比如关键词推荐、网页提取等等。

参 考 文 献

[许家梁，2009] 许家梁，*信息检索*，北京：国防工业出版社，2009

[王知津等，2005] 王知津，贾福新， 郑红军，*现代信息检索*，北京：机械工业出版社，2005

[朱彤，2010] 朱彤，*搜索引擎广告—网络营销的成功之路*，北京：电子工业出版社，2010

[于奎，2004] 于奎，*网络广告效果评价研究*，江西财经大学学报，2004

[李东，2005] 李东，*网络广告与传统广告的比较研究*，当代传播，2005

[杨小平等，2003] 杨小平，丁洁，黄都培，*基于向量空间模型的中文信息检索技术研究*，计算机工程与应用，2003

[李师贤等，2003] 李师贤，明仲，*大规模 C++程序设计*，北京：中国电力出版社，2003

[李建中 等，2006] 李建中，张岩，李治军，*数据结构（C 语言版）*，北京：机械工业出版社，2006

[宗成庆，2008] 宗成庆，*统计自然语言处理*，北京：清华大学出版社，2008

[苑春法，2005] 苑春法，李庆中，王昀，李伟，曹德芳，*统计自然语言处理基础*，北京：电子工业出版社，2005

[朱胜 ，2009] 朱胜 ，*互联网时代统计数据的搜集与分析方法*，北京：中国统计出版社，2009

[张友生，2009] 张友生，王勇，*系统架构设计师教程*，2009

[ccw，2011] http://news.ccw.com.cn/topic/20090325ads/，互联网广告调查，2011

[sina，2007] http://tech.sina.com.cn/focus/2007_GUIDE2008/，中国互联网调查报告，2007

[百科，2011] http://baike.baidu.com/view/136475.htm，JSON，2011

[baidu，2011] http://wm.baidu.com/，百度网盟推广，2011

[IBM，2008] http://www.ibm.com/developerworks/cn/web/wa-lo-json JSON 入门指南，2008

[google，2011] http://www.google.com.hk/intl/zh-CN/ads/，google 广告解决方案，2011

[Bruce ，2009] Bruce Croft，Donald Metzler ，Trevor Strohman，*Search Engines-Information Retrieval in Practice*，2009

[Harold，2006] Harold Davis ，*google Advertising Tools: Cashing in with AdSense, AdWords, and the Google APIs* ， New York:O'Reilly Media, Inc.，2006

[Wiki-PLSA，2011] http://en.wikipedia.org/wiki/Probabilistic_latent_semantic_analysis，Probabilistic latent semantic analysis，2009

[Wiki-LDA，2011] http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation，Latent Dirichlet allocation，2011

[Thomas，1999] Thomas Hofmann, *Probabilistic Latent Semantic Indexing*, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999

致　　谢

衷心感谢我的指导老师，感谢郑滔老师和刘钦老师在这近一年的论文写作过程当中给予我的悉心指导和帮助。郑老师和刘老师平易近人的生活作风、严谨求实的治学态度以及一丝不苟、勤勤恳恳的工作精神将永远是我学习的榜样。

感谢百度 ECOM 部门给我提供一个良好的学习和实践的机会，一个能将理论知识运用于实践的理想场所。在这里，有幸参与到关于广告检索相关工作，让我接触到了很多业界最新的技术和观念，获得了许多书本上学不到的知识和技能，这将对我以后的生活和工作产生深远的影响。感谢项目组的同事们在项目的开发过程中给我的无私帮助和指导，从他们身上我学到了很多工作的方法和做人的道理，在百度实习的日子是我永远难忘的记忆。

能够成为南京大学软件学院的硕士学员，能够在这里结识这位老师和朋友，我感到非常骄傲！衷心祝愿南京大学软件学院能够越来越灿烂和辉煌。

最后，向答辩委员会的各位老师致以深深的谢意，感谢你们所付出的辛勤的工作！

版权及论文原创性说明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含其他个人或集体已经发表或撰写的作品成果。本文所引用的重要文献，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名：

日期：　　年　　月　　日