预测网络新闻受欢迎程度的主动式智能决策支持系统

凯尔文-费尔南德斯1(✉), 佩德罗-维纳格2, 和保罗-科特兹2

[1]INESC TEC Porto/Universidade Do Porto, Porto, 葡萄牙
[2]ALGORITMI研究中心，米尼奥大学，布拉加，葡萄牙
kelwinfc@gmail.com

摘要。由于网络的扩张，网上新闻流行度的预测正在成为一个时尚的研究话题。在本文中，我们提出了一个新颖的、主动的智能决策支持系统（IDSS），在文章发表前对其进行分析。使用一组广泛的提取特征（例如，关键词、数字媒体内容、文章中引用的新闻的早期流行度），IDSS首先预测一篇文章是否会变得流行。然后，它优化了一个更容易被作者改变的文章特征子集，寻找预测的流行概率的提升。使用最近收集的一个大型数据集，包括来自Mashable网站的39,000篇文章，我们对五个最新的模型进行了稳健的滚动窗口评估。最好的结果是由随机森林提供的，辨别力为73%。此外，我们还探索了几种随机爬坡的局部搜索。当优化1000篇文章时，最好的优化方法在估计流行概率方面获得了15个百分点的平均增益。这些结果证明了所提出的IDSS是一个对在线新闻作者有价值的工具。

关键词：人气预测 - 在线新闻 - 文本挖掘 - 分类 - 随机局部搜索

1 简介

决策支持系统（DSS）是在1960年代中期提出的，涉及使用信息技术来支持决策。由于该领域的进步（如数据挖掘、元启发法），人们对智能DSS（IDSS）的发展越来越感兴趣，该系统采用人工智能技术进行决策支持[1]。自适应商业智能（ABI）的概念是2006年提出的一种特殊IDSS[2]。ABI系统将预测和优化结合在一起，而IDSS通常将两者分开处理，以便更有效地支持决策。其目标是首先使用数据驱动的模型来预测未来更有可能发生的事情，然后使用现代优化方法，在当前可知和可预测的情况下寻找最佳解决方案。

F.Pereira et al. (Eds.) EPIA 2015, LNAI 9273, pp.535-546, 2015.
doi: 10.1007/978-3-319-23485-4 53

在互联网和Web 2.0的扩展过程中，人们对在线新闻的兴趣也越来越浓厚，这使得信息可以在全球范围内轻松快速地传播。因此，预测在线新闻的受欢迎程度正成为最近的研究趋势（例如，[3,4,5,6,7]）。流行度通常是通过考虑网络和社交网络中的互动数量（例如，分享、喜欢和评论的数量）来衡量。预测这种流行度对作者、内容提供商、广告商甚至活动家/政治家（例如，了解或影响公众舆论）都很有价值[4]。根据Tatar等人的研究[8]，有两种主要的人气预测方法：一种是使用发布后才知道的特征，另一种是不使用这种特征。第一种方法更常见（例如，[3,5,9,6,7]）。由于预测任务比较容易，通常可以达到较高的预测准确率。后一种方法更加稀缺，虽然可能会出现较低的预测性能，但预测结果更加有用，允许（如在本工作中进行的）在出版前改进内容。

使用第二种方法，Petrovic等人[10]使用与推文内容相关的特征（如哈希标签的数量、提及、URL、长度、字数）和与作者相关的社会特征（如关注者的数量、朋友、用户是否经过验证）预测转发的数量。2010年10月期间共检索了2100万条推文。使用二元任务来区分被转发和未被转发的帖子，当同时使用推文内容和社会特征时，最高的F-1得分达到47%。同样，Bandari等人[4]关注四种类型的特征（新闻来源、文章的类别、使用的副标题语言和文章中提到的名字）来预测提到一篇文章的推文数量。该数据集从Feedzilla获取，并与一周的数据相关。测试了四种分类方法来预测三个流行等级（1到20条推文，20到100条推文，100条以上；没有推文的文章被丢弃），结果是Na¨ve Bayes和Bagging的准确率分别为77%到84%。最后，Hensinger等人[11]测试了两个预测二元分类任务：流行/不流行和吸引/不吸引，当与同一天发表的其他文章相比。该数据与10个英语新闻机构有关，与一年有关。使用文本特征（如标题和描述的词包、关键词）和其他特征（如发表日期），结合支持向量机（SVM），作者在与流行/不流行任务相比时，对吸引人的任务获得了更好的结果，前者的准确率为62%到86%，后者为51%到62%。

在本文中，我们提出了一种新的主动式IDSS，在网上新闻发布*之前*对其进行分析。假设采用ABI方法，首先用预测模块估计一篇文章的受欢迎程度，然后用优化模块建议改变文章的内容和结构，以使其预期受欢迎程度最大化。在我们的知识范围内，以前没有任何工作涉及这种主动的ABI方法，结合预测和优化来改善新闻内容。预测模块使用了大量的输入，

包括纯粹的新特征（与文献[4,11,10]相比）：数字媒体内容（如图像、视频）；早期的流行------。

文章中引用的新闻的数量；发表前关键词的平均分享次数；以及自然语言特征（例如，标题的极性，Latent Dirich- let Allocation主题）。我们采用常见的二进制（流行/不流行）任务，并在现实的滚动窗口下测试五种最先进的方法（如随机森林、自适应提升、SVM）。此外，我们使用了时尚的Mashable（mashable.com/）新闻内容，这在以前预测流行度时没有被研究过，并收集了最近两年的大型数据集（与文献相比，这个时间段大得多）。此外，我们还使用局部搜索方法（随机爬坡）来优化新闻内容，该方法在部分特征集合中寻找增强点，从而使用户更容易改变。

2 材料和方法

2.1 数据采集和准备

我们从最大的新闻网站之一Mashable检索了过去两年中发表的所有文章的内容。这项工作中描述的所有数据收集和处理程序（包括预测和操作模块）都由作者用Python实现。数据是在2013年1月7日至2015年1月7日的两年时间内收集的。我们放弃了一小部分不遵循一般HTML结构的特殊场合文章，因为处理每一种场合类型都需要一个特定的解析器。我们也舍弃了最近的文章（少于3周），因为Mashable分享的数量对于其中一些文章来说没有达到收敛的程度（例如，少于4天），而且我们也希望在我们的滚动窗口评估策略中保持每个测试集的文章数量不变（见2.3节）。经过这样的预处理，我们最终得到了总共39,000篇文章，如表1所示。所收集的数据被捐赠给UCI机器学习库（http://archive.ics.uci.edu/ml/）。

表1. Mashable数据集的统计措施。

		每天的文章			
文章数量	**总天数**	**平均值**	**标准偏差**	**闵行区**	**最大**
39,000	709	55.00	22.65	12	105

我们从HTML代码中提取了大量的特征（共47个），以使这些数据适合学习模型，如表2所示。在该表中，属性类型被分为：数字--整数值；比率--在[0，1]内；布尔--∈{0，1}；以及名义。列**类型**在括号（#）内显示与属性相关的变量数量。与之相似的是

在[6,7]中所执行的，我们进行了对数转换，以缩放无限制的数字特征（例如，文章中的字数），而名义属性则用常见的*1-C*编码进行转换。

我们选择了一大串描述文章不同方面的特征，这些特征被认为可能与影响分享数量有关。其中一些特征取决于Mashable服务的特殊性：文章经常引用在同一服务中发表的其他文章；文章有元数据，如关键词、数据渠道类型和总分享数（当考虑Facebook、Twitter、Google+、LinkedIn、Stumble-Upon和Pinterest时）。因此，我们提取了文章中引用的所有Mashable链接的最低、平均和最高分享次数（发表前已知）。同样地，我们对所有文章关键词的平均分享量（发表前已知）进行排名，以获得最差、平均和最好的关键词。对于每一个关键词，我们提取最小、平均和最大的分享数。数据渠道的类别是："生活方式"、"公共汽车"、"娱乐"、"社会医学"、"科技"、"病毒 "和 "世界"。

我们还提取了几个自然语言处理特征。Latent Dirichlet Allocation（LDA）[12]算法被应用于所有Mashable文本（发表前已知），以便首先确定五个最相关的主题，然后衡量当前文章与这些主题的接近程度。为了计算主观性和极性情绪分析，我们采用了Pattern web mining模块（http://www.clips.ua.ac.be/pattern）[13]，允许计算情绪极性和主观性分数。

表2.按类别划分的属性列表。

特点	**类型**(#)
词条	
标题中的字数	数字 (1)
文章中的字数	数字 (1)
平均字长	数字 (1)
不间断字数的比率	比率 (1)
独特字数的比率	比率 (1)
独特的不间断词的比率	比例 (1)
链接	
链接的数量	数字 (1)
Mashable文章链接的数量	数字 (1)
最低、平均和最高数量的股份的Mashable链接	数字 (3)
数字媒体	
图像的数量	数字 (1)
视频的数量	数字 (1)
时间	
一周中的一天	名义上的 (1)
在周末出版？	

特点	**类型**(#)
关键词	
关键字的数量	数字 (1)
最差关键词（最低/平均/最高份额）	数字 (3)
平均关键词（最低/平均/最高份额）	数字 (3)
最佳关键词（最低/平均/最高份额）	数字 (3)
文章类别（Mashable数据频道）	名义上的 (1)
自然语言处理	
与前5个LDA主题的接近程度	比率 (5)
标题主观性	比例 (1)
文章文本的主观性得分及其与0.5的绝对差异	比率 (2)
标题情感的极性	比例 (1)
正面和负面词汇的比率	比率 (2)
正面词在非中性词中的比率	比例 (1)
负面词在非中性词中的比率	比例 (1)
正面词的极性（最小/平均/最大）	比率 (3)
负面词的极性（最小/平均/最大）	比率 (3)
文章文本的极性得分和其绝对差异为0.5	比率 (3)

目标	**类型**(#)
文章Mashable分享的数量	数字 (1)

2.2 智能决策支持系统

按照ABI的概念，拟议的IDSS包含三个主要模块（图1）：数据提取和处理、预测和优化。第一个模块执行第2.1节中描述的步骤，它负责

用于收集在线文章并计算其各自的特征。预测模块首先接收处理过的数据，并将其分割成训练、验证和测试集（数据分离）。然后，它对分类模型进行调整和拟合（模型训练和选择）。接下来，最好的分类模型被储存起来，用于提供文章的成功预测（流行度估计）。最后，优化模块搜索当前文章内容特征的子集的更好组合。在这个搜索过程中，会大量使用分类模型（神谕）。此外，一些新的搜索特征组合可能需要重新计算各自的特征（例如，平均关键词的最低分享数量）。在图中，这种依赖关系由特征提取和优化之间的箭头表示。一旦优化完成，将向用户提供一份文章修改建议的清单，让她/他做出决定。

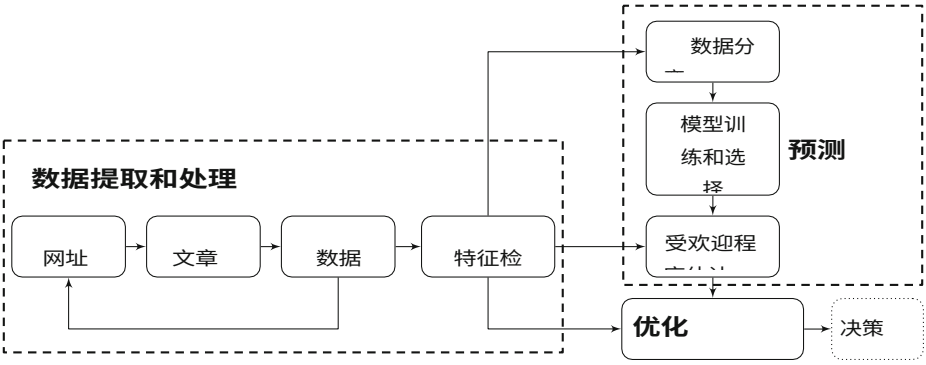

图1.描述IDSS行为的流程图。

2.3 预测模块

我们采用Scikit learn[14]库来拟合预测模型。与在[10,4,11]中执行的类似，我们假设一个二元分类任务、

其中，如果一篇文章的分享数量高于一个固定的决策阈值（D_1），则被视为 "受欢迎"，否则被视为 "不受欢迎"。

在本文中，我们测试了五个分类模型：随机森林（RF）；自适应提升（AdaBoost）；带有径向基函数（RBF）角的SVM；K-近邻（KNN）和Na¨ve Bayes（NB）。使用网格搜索来寻找以下的最佳超参数：RF和AdaBoost（树的数量）；SVM（*C*权衡参数）；和KNN（邻居的数量）。在这个网格搜索过程中，训练数据在内部被分成训练集（70%）和验证集（30%），方法是使用随机保留分割。一旦选择了最佳的超参数，那么该模型就会被拟合到所有的训练数据上。

接受者操作特征（ROC）曲线显示了两类分类器在可能的阈值（$D_2 \in [0, 1]$）范围内的性能，绘制了特异性（*X轴*）与敏感性（*Y轴*）的减一图[15]。在
这项工作中，分类方法假设了一个概率模型，如果一个类别的预测概率为$p>D_2$，则被认为是积极的。我们计算了几个分类指标：准确率、精确率、召回率、F1得分（均使用固定的$D_2 = 0.5$）；以及ROC下的面积（AUC，它考虑到所有的
D_2 值）。AUC指标是最相关的指标，因为它衡量的是
分类器的识别能力，它与所选的D_2 值无关[15]。理想的方法应该呈现1.0的AUC，而AUC为0.5
表示一个随机分类器。为了实现稳健的评价，我们采用了滚动窗口分析法[16]。在这种评估方式下，一个由W个连续样本组成的训练窗口被用来拟合模型，然后进行L次预测。接下来，训练窗口被更新，用最近的L个样本代替最旧的L个样本，以适应新的模型并进行新的L个预测，依次类推。

2.4 优化

本地搜索通过在初始解决方案的附近搜索来优化一个目标。这种类型的搜索适合我们的IDSS优化模块，因为它接收一篇文章（初始解决方案），然后试图通过搜索可能的文章变化（在初始解决方案的附近）来增加其预测的流行概率。一个简单的局部搜索方法的例子是爬坡法，它在当前解决方案的附近反复搜索，并在发现更好的解决方案时更新该解决方案，直到达到局部最优或停止该方法。在本文中，我们使用了随机爬坡法[2]，其工作原理与纯爬坡法相同，只是最差的解决方案可以以P的概率被选中。 我们测试了几个P值，范围从P=0（爬坡法）到P=1（蒙特卡洛随机搜索）。

为了评估解决方案的质量，局部搜索使 "流行 "类的概率最大化，这是由最佳分类模型提供的。此外，搜索只在更适合由作者改变的特征子集上进行（内容的调整或出版日的改变），详见表3。在每次迭代中，邻域搜索空间假定特征原始值有小的扰动（增加或减少）。例如，如果当前标题中的字数为n=5，那么就会执行较短 （n' =4）或较长 （n' =6）的标题搜索。由于星期被表示为一个名义变量，在扰动中假定随机选择一个不同的日子。同样地，鉴于关键词集 （K） 不是数字，我们提出了一个不同的扰动策略。对于一篇特定的文章，我们计算出一个建议的关键词列表K'，其中包括在文本中出现过一次以上的、在以前的文章中被用作关键词的词。为了保持问

题的可计算性，我们只考虑了在以前的平均份额方面最好的五个关键词。然后，我们通过增加一个建议的关键词或删除一个建议的关键词来产生扰动。

的原始关键词。优化N篇文章（即N个本地搜索）时的平均性能，使用平均增益（MG）和转换率（CR）进行评估：

$$MG = \frac{1}{N}\sum_{i=1}^{N}(Q_i - Q)_i \qquad (1)$$

$$CR = \overline{U'/U}$$

其中，Q_i表示原始文章（i）的质量（估计流行概率），Q'是使用本地搜索获得的质量，U是不受欢迎的文章的数量（估计概率≤D_2，对于所有N篇原始文章）和
U'是转换后的文章数量（原来估计的概率是≤D_2
但优化后改为> D_2）。

表3.可优化的功能。

特点	**扰动**
标题中的字数(n)	$n' \in \{n-1, n+1\}, n \geq 0 \wedge n' \neq n$
内容中的字数（n）。	$n' \in \{n-1, n+1\}, n \geq 0 \wedge n' \neq n$
图像的数量（n）。	$n' \in \{n-1, n+1\}, n \geq 0 \wedge n' \neq n$
视频的数量(n)	$n' \in \{n-1, n+1\}, n \geq 0 \wedge n' \neq n$
一周中的一天(w)	$w' \in [0..7], w' \neq w$
关键词 (K)	$k' \in \{K \cup i\} \cup \{K - j\}, i \in K' \boxtimes j \in K$

3 实验和结果

3.1 预测

对于预测实验，我们采用了滚动窗口方案，训练窗口大小为W=10,000，在每个迭代中进行L=1,000次 预测。在这种设置下，每个分类模型被训练了29次（迭代），产生了29个预测集（每个大小为L）。对于定义一个受欢迎的类别，我们使用一个固定值D_1 = 1, 400股，这导致了在第一个训练集中 "受欢迎的"/"不受欢迎的 "类别分布平衡。

(前10, 000篇)。选择的超参数的网格搜索范围是：RF和AdaBoost - 树的数量∈{10, 20, 50, 100, 200, 400}；SVM - $C \in \{2^0, 2^1, \ldots, 2^6\}$；而KNN--邻居的数量∈{1，3，5，10，20}。

表4显示了所获得的分类指标，这些指标是通过对联盟的计算得出的。所有29个测试集的结果。在表格中，根据AUC指标对模型的性能进行了排名。图2的左边是最佳（RF）、最差（NB）和基线（对角线，对应于预测）模型的ROC曲线。该图证实了RF比NB的优越性。

所有D_2阈值的模型，包括更敏感（X轴值接近零，$D_2 >> 0.5$）或特定（X

*轴*接近一，$D_2 << 0.5$）的权衡。对于最佳模型（RF），图2的右图显示了AUC指标的演变情况

在滚动窗口迭代过程中，揭示了一个有趣的稳定的预测性能，随着时间的推移。获得的最佳结果（AUC=0.73）比随机分类器高23个百分点。虽然不是很完美，但却达到了有趣的消除水平，高于70%。

表4.滚动窗口评价的模型比较（最佳值为**黑体**）。

模型	**准确度**	**精度**	**召回率**	F1	AUC
随机森林 (RF)	**0.67**	0.67	**0.71**	**0.69**	**0.73**
自适应提升（AdaBoost）。	0.66	0.68	0.67	0.67	0.72
支持向量机(SVM)	0.66	0.67	0.68	0.68	0.71
K-Nearest Neighbors (KNN)	0.62	0.66	0.55	0.60	0.67
Na¨ve Bayes (NB)	0.62	**0.68**	0.49	0.57	0.65

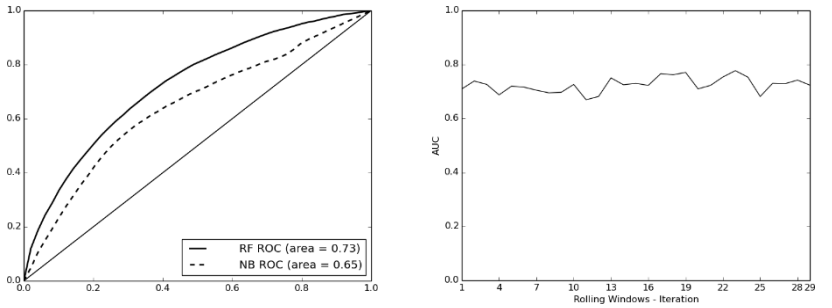

图2.RF的ROC曲线（左）和AUC指标在t ime的分布（右）。

表5显示了相对重要性（列**Rank**显示比率值，#表示特征的排名），由RF算法在用所有数据（39,000篇文章）训练时测量。由于空间的限制，该表显示了最好的15个特征，也是优化模块所使用的特征。关键词相关的特征具有更强的重要性，其次是基于LDA的特征和Mashable链接的份额。特别是，在下一节中被优化的特征（**有关键词**子集）在RF模型中具有很强的重要性（33%）。

3.2 优化

在优化实验中，我们使用了最佳分类模型（RF），这是在滚动窗口方案的最后一次迭代中训练出来的。然后，我们从最后的测试集（N=1,000）中选择所有的文章来评估本地的
搜索方法。我们测试了六个随机爬坡概率（$P\in$）。

{0.0, 0.2, 0.4, 0.6, 0.8, 1.0}).我们还测试了两个特征优化子集

表5.根据其在RF模型中的重要性对特征进行排名。

特点	排名(#)	特点	排名(#)
平均关键词(平均股数)	0.0456 (1)	与前1名LDA主题的接近程度	0.0287 (11)
平均关键词（最大份额）	0.0389 (2)	独特的不间断词的比率	0.0274 (12)
与前3名LDA主题的接近程度	0.0323 (3)	文章文本的主观性	0.0271 (13)
文章类别（Mashable数据频道）	0.0304 (4)	独特标记词的比率	0.0271 (14)
Mashable链接的最小份额	0.0297 (5)	平均令牌长度	0.0271 (15)
最佳关键词（平均份额）	0.0294 (6)	字数	0.0263 (16)
Mashable链接的平均份额	0.0294 (7)	一周中的一天	0.0260 (18)
与前2个LDA主题的接近程度	0.0293 (8)	标题中的字数	0.0161 (31)
最差的关键词（平均份额）	0.0292 (9)	图像的数量	0.0142 (34)
与前5名LDA主题的接近程度	0.0288 (10)	视频的数量	0.0082 (44)

与表3相关：使用除关键词外的所有特征（*无关键词*）和使用所有特征（*有关键词*）。每个局部搜索在100次迭代后停止。在搜索过程中，我们存储与迭代相关的最佳结果

$I \in \{0, 1, 2, 4, 8, 10, 20, 40, 60, 80, 100\}$.

图3显示了对以下情况的最终优化性能（经过100次迭代）：1.在考虑两个特征扰动子集时，随机概率参数P的变化。图3还显示了局部搜索的收敛情况（对于不同的P值）。P的极端值（0-纯爬坡；1-随机搜索）与它们的邻近值相比，产生了较低的性能。特别是，图4显示，纯爬坡法过于贪婪，执行了一个快速的初始收敛，但很快就变得平淡。当使用*无关键词子*集时，P的最佳值是MG的0.2和CR指标的0.4。对于*有关键词子*集，两个优化指标的最佳P值都是0.8。此外，纳入关键词相关的建议对优化产生了很大的影响，提高了两个指标的性能。例如，在最佳情况下，MG指标从0.05增加到0.16（$P = 0.8$）。此外，图3显示，与*有关键词*搜索相比，*无关键词子*集优化是一项更容易的任务。正如Zhang和Dimitroff[17]所认为的，元数据对网页的可见度有重要作用，这可能解释了关键词在预测（表5）和优化流行度（图3）时的重要影响。

出于演示的目的，图5显示了已实现的IDSS原型的界面示例。一篇较新的文章（来自2015年1月16日）被选来做这个演示。在这种情况下，IDSS使用*无关键词子*集，如果执行几个变化，例如将标题词的数量从11个减少到10个，估计流行概率会增加13个百分点。在另一个例子中（图中未显示），使用*有关键词子*集，IDSS建议从以下几个方面进行改变

将关键词K⍰{"电视"、"节目时间"、"未分类"、"娱乐"、"电影"、"国土"、"回顾"}添加到$K' \in$⍰{"电影"、"关系"、"家庭 "和 "夜晚"}中，作为一篇关于《国土》电视节目结束的文章。

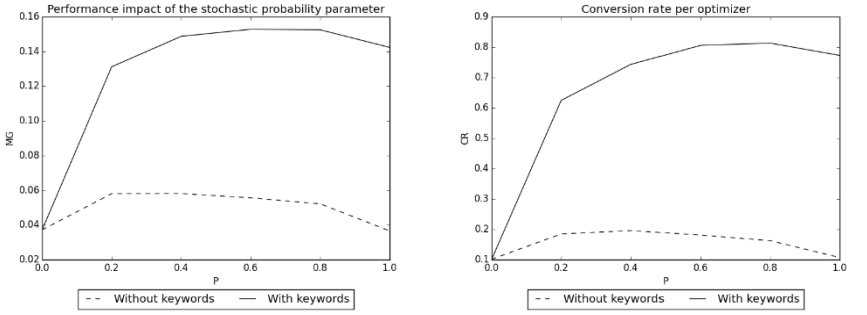

图3.随机概率（*P*）对平均增益（左） 和转换率（右）的影响。

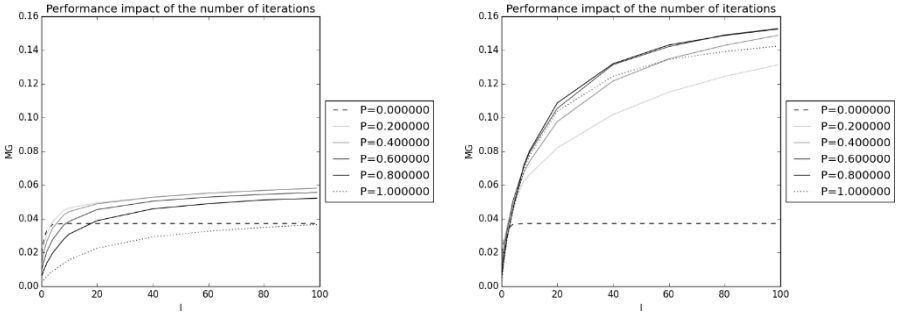

图4.在*无关键词 ds*（左）和 *有关键词*（右）特征子集下的局部搜索收敛情况（*y 轴*表示平均增益，*x 轴*表示迭代次数）。

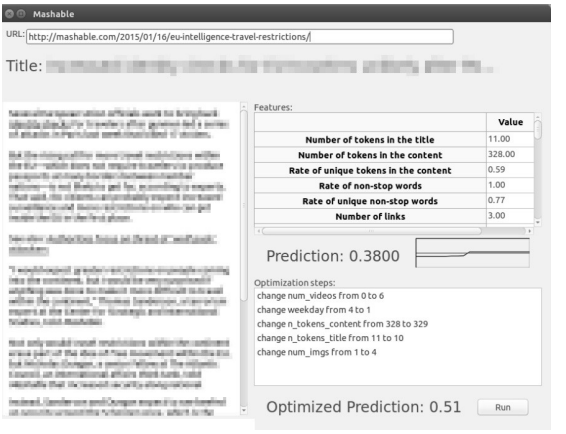

图5. IDSS原型的界面实例。

4 结论

随着网络的扩展，人们对预测在线新闻的受欢迎程度越来越感兴趣。在这项工作中，我们提出了一个智能决策支持系统（IDSS），它首先提取了文章发表前已知的广泛的特征集，以便在二元分类任务下预测其未来的受欢迎程度。然后，它优化了文章特征的一个子集（更适合由作者改变），以提高其预期的受欢迎程度。

我们使用大型的最新数据集，从流行的Mashable新闻服务中收集了39,000篇文章，并进行了滚动式赢余评估，在不同的指标下测试了五个最新的分类模型。总的来说，最好的结果是由随机森林（RF）实现的，接收器操作特征（ROC）曲线下的总面积为73%，这相当于一个可接受的区分。我们还分析了RF输入的重要性，发现基于关键词的特征是最重要的特征之一，其次是自然语言处理特征和Mashable链接的先前份额。使用最佳预测模型作为神谕，我们探索了几种随机爬坡搜索的变体，目的是在改变文章特征的两个子集（例如，标题中的字数）时增加估计的文章概率。当优化1,000篇文章（来自最后的滚动窗口测试集）时，我们在最佳局部搜索设置的平均增益方面取得了15个百分点。考虑到所获得的结果，我们相信所提出的IDSS对于Mashable的作者来说是相当有价值的。

在未来的工作中，我们打算探索更多的与con- tent有关的高级功能，如趋势分析。此外，我们还计划对文章进行长期跟踪，以便使用更复杂的预测方法。

鸣谢。这项工作得到了FCT - Funda¸c˜ao para a Cˆencia e Tecnologia的支持，项目范围是UID/CEC/00319/2013。作者要感谢Pedro Sernadela在之前工作中的贡献。

参考文献

1. Arnott, D., Pervan, G.: 决策支持系统学科的八个关键问题。决策支持系统44(3), 657-672 (2008)
2. Michalewicz, Z., Schmidt, M., Michalewicz, M., Chiriac, C. ：Adaptive business intelligence.Springer (2006)
3. Ahmed, M., Spagna, S., Huici, F., Niccolini, S.: 窥视未来：预测用户生成内容的流行演变。在：第六届ACM网络搜索和数据挖掘国际会议论文集，第607-616页。ACM (2013)
4. Bandari, R., Asur, S., Huberman, B.A.: The pulse of news in social media: forecasting popularity.在：ICWSM (2012)

5. Kaltenbrunner, A., Gomez, V., Lopez, V.: Slashdot活动的描述和预测。In：网络会议，LA-WEB 2007，第57-66页。IEEE，拉丁美洲(2007)

6. Szabo, G., Huberman, B.A.: 预测在线内容的流行。Commu- nications of the ACM 53(8), 80-88 (2010)
7. Tatar, A., Antoniadis, P., De Amorim, M.D., Fdida, S.: 从人气预测 到在线新闻排名。社会网络分析和挖掘4(1), 1-12 (2014)
8. Tatar, A., de Amorim, M.D., Fdida, S., Antoniadis, P.：关于预测网络内容流行度的调查。互联网服务与应用杂志》5(1), 1-20 (2014)
9. Lee, J.G., Moon, S., Salamatian, K.: 用考克斯比例危险回归模型对在线内容的受欢迎程度进行建模和预测。Neurocomputing 76(1), 134-145 (2012)
10. Petrovic, S., Osborne, M., Lavrenko, V.: RT赢！预测Twitter中的消息传播。In：第五届AAAI网络日志和社会媒体国际会议 (ICWSM)，第586-589页(2011)
11. Hensinger, E., Flaounas, I., Cristianini, N.: 建模和预测新闻 流行。模式分析与应用》16(4), 623-635 (2013)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation.机器杂志 学习研究 3, 993-1022 (2003)
13. De Smedt, T., Nijs, L., Daelemans, W.: Creative Web services with pattern.在：第五届计算创造力国际会议论文集 (2014)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. ：Scikit-learn：Python中的机器 学习。机器学习研究杂志》12, 2825-2830 (2011)
15. Fawcett, T.: An introduction to roc analysis.Pattern Recognition Letters 27(8), 861-874 (2006)
16. Tashman, L.J.：预测准确性的样本外测试：分析和回顾。International Journal of Forecasting 16(4), 437-450 (2000)
17. Zhang, J., Dimitroff, A.: 元数据实施对网页在搜索引擎结果中可见度的影响（第二部分）。信息处理与管理41(3), 691-715 (2005)