IFLA WLIC 2013
SINGAPORE

Submitted on: July 22, 2013

利用现有的半自动分类系统开发基于本体的全自动文档分类系统

Translation of the original paper "An Ontology Based Fully Automatic Document Classification System Using an Existing Semi-Automatic System"

Chaaminda Manjula Wijewickrema
Main Library, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka.
manju@sab.ac.lk

Ruwan Gamage
National Institute of Library and Information Sciences, University of Colombo, Colombo, Sri Lanka.
mailruga@gmail.com

Translated by <张士男，Shinan Zhang >, <中国科学院国家科学图书馆，National Science Library, Chinese Academy of Sciences>, <中国，China>

摘要:

由于数字内容的指数增长和手动组织、半自动组织的非高效性，文档自动分类已经成为一个重要的研究领域。一方面，手动和半自动分类需耗费大量精力并且是劳动密集型，另一方面，这两种方法中由于文档的模糊性和分类表所带来的误分类不可避免。

因此，本研究试图解决这些问题。本研究提出一个自动化系统，这个自动化系统完全可以通过最小化词汇歧义为一个给定的文本文档进行分类。我们前期已经开发了一个半自动文档分类系统，这里对其进一步优化以获得一个全自动的文档分类系统。

关键词:自动分类， 文本分类， 本体， tf-idf 权重函数

1 引言

近年来，可用的印刷版和电子版的信息量急剧增加。现代世界中，这些迅速增长的信息导致人们用易于访问的方式组织文本材料，同时，标准的图书分类法用来实现这一目标。然而，自然语言固有的一些缺陷可能会在分类过程中产生严重问题，特别是自然语言中存在的词汇模糊性增加了分类不准确性。例如，同音异义词“ ontology”，计算机科学中的“ontology ”可能被归入哲学中的“ontology ”。反过来，这会影响一个文档被发现的可能性。在传统分类方法中用于组织材料花费的时间和精力也很高，此外，数字化文档的体量迅速增加，要求更加简单和经济实惠的机械化方法来进行组织。

文本自动分类（ATC）是一个关于从数字文档自动分类到预定义类的研究领域。事实上，分类算法或文本分类器被称为是影响 ATC 的主要因素。支持向量机，概率方法，泛型算法，远程学习法，隐马尔可夫模型，决策树法，回归法，决策规则法，神经网络法（Sebastiani, 2002; Tao, Ling, & Cheng, 2005）和基于的 tf-idf （term frequency-inverse document frequency）（Abbas, Smaili, & Berkani, 2010; Tao, Ling, & Cheng, 2005）是文本分类时经常使用的分类算法。

本研究中，我们试图进一步优化(Wijewickrema & Gamage, 2012a, 2012b)介绍的基于 tf-idf 的半自动（混合）文本分类系统（本文称之为 HTCS）。在这里，进一步优化了 tf-idf 的基础功能，使用了领域本体来减少词汇歧义性。

尽管 HTCS 比手动方法结果更佳，但由于其分类过程的半自动性而存在一定局限。由于人工干预在分类过程中没有完全消除，最终的分类仍然依赖于人的决定。一方面，半自动分类系统选出几个候选分类，分类人员不得不在其中选出最合适的分类。另一方面，全自动方法能够进一步减少花费在特定任务上的时间和精力。

2 文献回顾

在语言学和自然语言处理领域对自然语言的歧义已经进行了广泛讨论(Richardson & Smeaton, 1995)。1985 年，普林斯顿大学启动了 WordNet 项目(Miller et al., 1990; Morato et al, 2004)来开发英语词汇资源。此外，还开发了许多其他电子词汇资源(Best, Nathan, & Lebiere, 2010; Prevot, Borgo, & Oltramari, 2005; Valitutti, Strapparava & Stock, 2004)。

有很多使用词汇资源（或本体）进行自动分类的尝试，Prabowo et al. (2002) 和 Song et al. (2005)报道了基于本体的网页分类系统，前者使用了基于国会图书馆分类法（LCC）和杜威十进制分类法（DDC）的本体。尽管如此，Song et al. (2005), Prabowo et al.’s (2002)的研究由于不适用于创建复杂分类而很少被使用。Song et al. (2005)用半自动化方法开发了经济学领域本体，但由于这个本体的可描述性差，人们不能得到期望的高度精确的分类结果。

除了网页资源分类，本体还应用于电子邮件、新闻等数字格式资源的分类。Taghva et al. (2003)制定了一个基于本体的电子邮件分类系统，Tenenboim, Shapira, and Shoval (2008)报道了一个基于本体的电子报纸分类系统。

3 方法

该系统主要基于三个关键的分类阶段。

3.1 第一个分类阶段

3.1.1 遏止和删除停用词

该系统第一阶段关注的是从索引术语到根术语的遏止和停用词的消除来减少从输入文档阶段起产生的无意义词汇。完成这两件事后的成品文件（目前仅限为一个词汇列表）被作为这一过程的输出，此输出中的前几个高频术语被选作为确定输入文档学科的关键术语。

3.1.2 训练集

训练集是一组被存储的文档的集合。人们可用他们和输入文档对比进而确定训练集的主题，可通过输入文档先前已选的关键术语和训练集文档中的术语进行相似度评估来实现。本研究使用了哲学领域的 385 个训练文档，这些文档已经按照经验丰富的学科分类器 DDC 进行了分类。

3.1.3 文本分类算法

一般来讲，文本分类算法在数量上决定了某一给定文档与某一特定主题的相关程度。之前的研究中(Wijewickrema & Gamage, 2012b)开发了一种文本分类器算法，这种算法使用现有术语频次权重函数即 tf-idf 权重函数开发(Salton & Buckley, 1988)。本研究使用了相同的分类器，对比输入文档和每一个训练文档的相似性。因此，获得最高数值的训练文档的学科被确定为输入文档的学科，新分类器使用的公式如下：

$$Document\ Score = \sum_{i=1}^{4} \frac{(tf-idf)_{i,D}}{\sqrt{\sum_{k=1}^{4}(tf-idf)_{k,D}^{2}}} \times (tf-idf)_{i,d} \qquad (1)$$

和 分别表达输入文档 D 和训练文档 d 中术语t_i的 tf-idf 权重。这里我们限定总和最多是四个，因为我们认为只有前四个高频术语才能决定输入文档的主题。

$(tf-idf)_{i,j}$ 定义如下：

$$(tf-idf)_{i,j} = \frac{f_{i,j}}{\sum_k f_{k,j}} \times \left[1 + \ln\left(\frac{|N|}{|n_i|+1}\right)\right] \qquad (2)$$

$f_{i,j}$ 表示文本文档d_j中关键词t_i的频次，N 是集合中全部文档的数量，n_i是关键词t_i出现的文档的编号。一般来讲，ti-idf 权重函数有能力基于几个因素为文档分配数据值，他们包括术语频次、文档中全部术语的总数、集合中出现某一特定词汇的文档数量和语料库中全部文档的数量。因此，tf-idf 权重函数可被认作是一个基本的分类器。这一基本形式存在一些缺点，例如只考虑了单个关键词来分类而不考虑与其相同频率的其他术语的重要性，新分类器能够在某种程度上克服这些缺点。

3.2 第二个分类阶段

第二阶段利用领域本体减少对分类中间过程造成影响的词汇歧义。我们开发了一个程序来将第一阶段的分类结果扩展至本体，因此，在分类的第二阶段获得了给定输入文档的候选分类。

3.2.1 本体

文本文档分类中的词汇消歧是本研究的主要研究对象。因此倾向于利用一个结构良好的本体来实现。为了建立本体，我们使用了两个资源，首先利用 DDC 进行本体创建，然后利用 Sears（希尔斯主题词表）进行语义丰富。本体限定在哲学和超自然现象相关主题领域，这些主题分布在 DDC 21 版中的 110 到 139 号段，进行主题限定的原因是构建大规模知识库的实际困难，因此本研究也倾向于对这一主题范围的文档进行分类。但术语和领域之间的语义映射方法同样适用于其他领域。进一步讲，特别需要注意的是一个新本体要创建如下关系：同义词、近义词、上位词、下位词。
例如，图 1 展示了术语‘cosmology’在本体中的呈现。

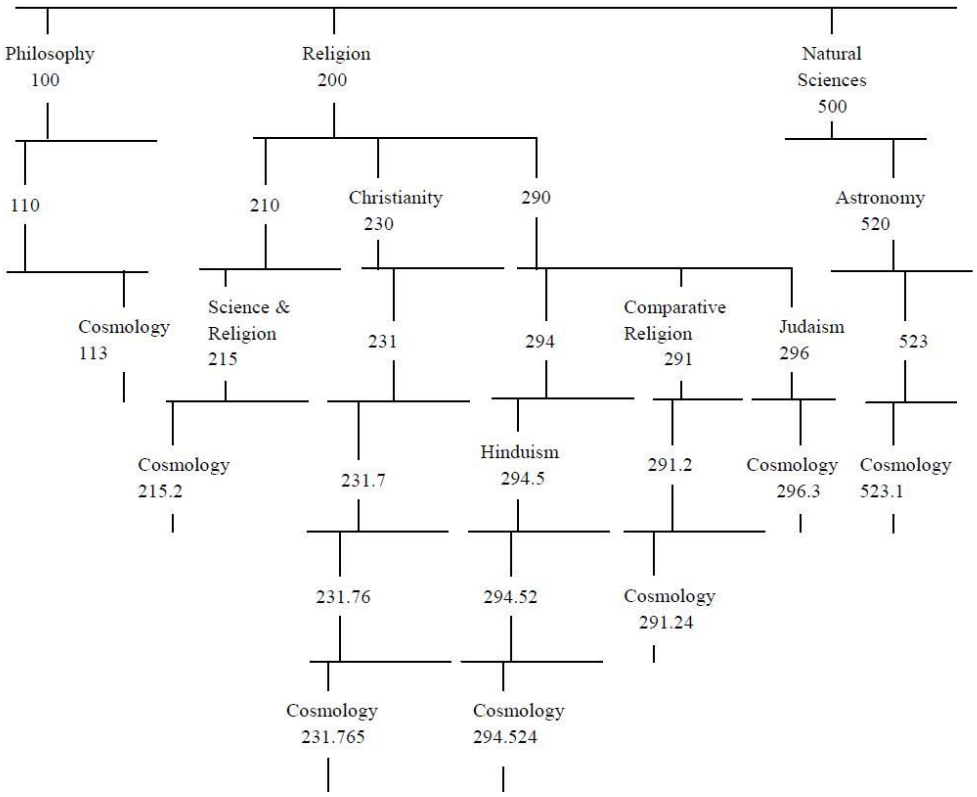

Figure 1: Structural arrangement of the term ‘cosmology’ in the domain ontology

如图 1 所示，术语‘cosmology’出现在了本体的多个位置，并且本体适当的创建了他们之间的关系。所以，一旦我们在本体中查询术语‘cosmology’，可能会获得如下检索结果：

Astronomy-Cosmology_InPhilosophyOfReligion_215.2
Cosmology_InAstronomy_523.1
Cosmology_InCreation_InChristianity_231.765
Cosmology_InHinduism_294.524
Cosmology_InPhilosophy_113
Creation-Cosmology_InReligion_291.24
Theology-Ethics-ViewsOfSocialIssues_InJudaism_296.3

3.3 第三个分类阶段

最后阶段采用过滤第二阶段产生的所有可能的分类结果这一策略，使用户只获得一个分类建议。如图 2 所示，将第二阶段给出的术语和原始输入文档中的术语进行对比过滤，这一过程中匹配程度最高的学科被认为是给定输入文档最合适的主题。
图 2 展示了分类过程的主要阶段。

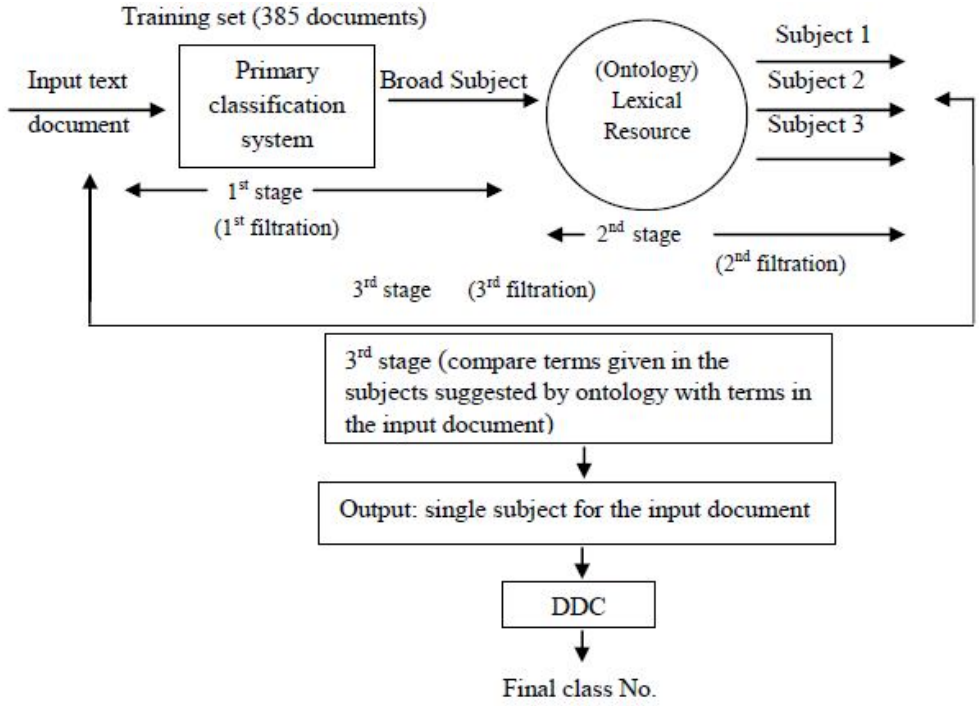

Figure 2: All stages of the fully automatic classification system.

3.4 用于实现系统的工具

第一阶段使用 Lucene API（Application Programming Interface）来选择适用于给定的输入文档最相关的主题。为了启动第二阶段，这些选定的主题标签被发送到一个预先设置的 OWL 本体中，这一本体利用 Protégé 本体编辑器创建。OWL 格式本体创建之后，利用 Protégé-OWL API 将其与 Lucene API 进行比较，这里， Protégé-OWL API 被当做是一个从 OWL 本体文件中检索信息的工具。作为第二阶段的输出，系统产生了几个候选类目。我们将这些输出作为第三阶段的输入来选出一个与输入文档最相关的主题，为了实现这一目的，再次使用了 Lucene API。

图 3 展示了每一个工具在整个系统中的使用。

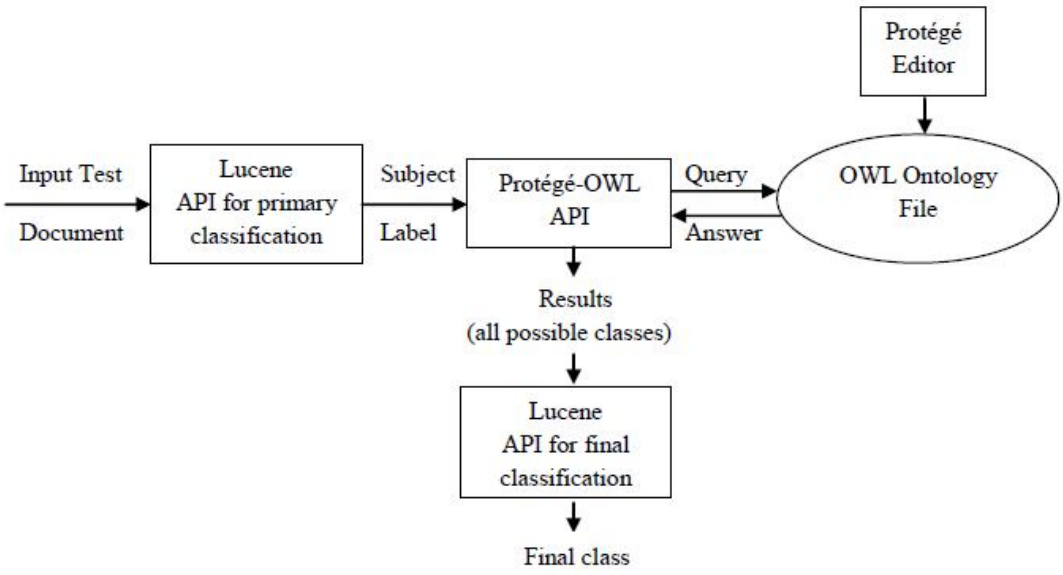

Figure 3: Implementing the system using tools

4 结论

在之前的研究中(Wijewickrema and Gamage, 2012a)，一个经验丰富的主题分类人员对 58 个输入文档按照 DDC 进行了分类。进一步讲，这个分类人员被要求阅读文档给出所有可能的主题/学科，并选出最合适的一个。然后，同一个主题分类人员借助半自动系统对相同的文档集合进行分类，这个分类人员重新在系统给出的所有分类中选出最合适的主题。按照这一顺序，在第二次过滤后，我们获得了利用手工和半自动方式得到的输入文档主题，由于输入文档进行了预分类，我们进行了结果准确性的比较。

例如，图 4 展示了第二步过滤后获得的分类结果，输入文档术语‘其他宗教’这一主题。

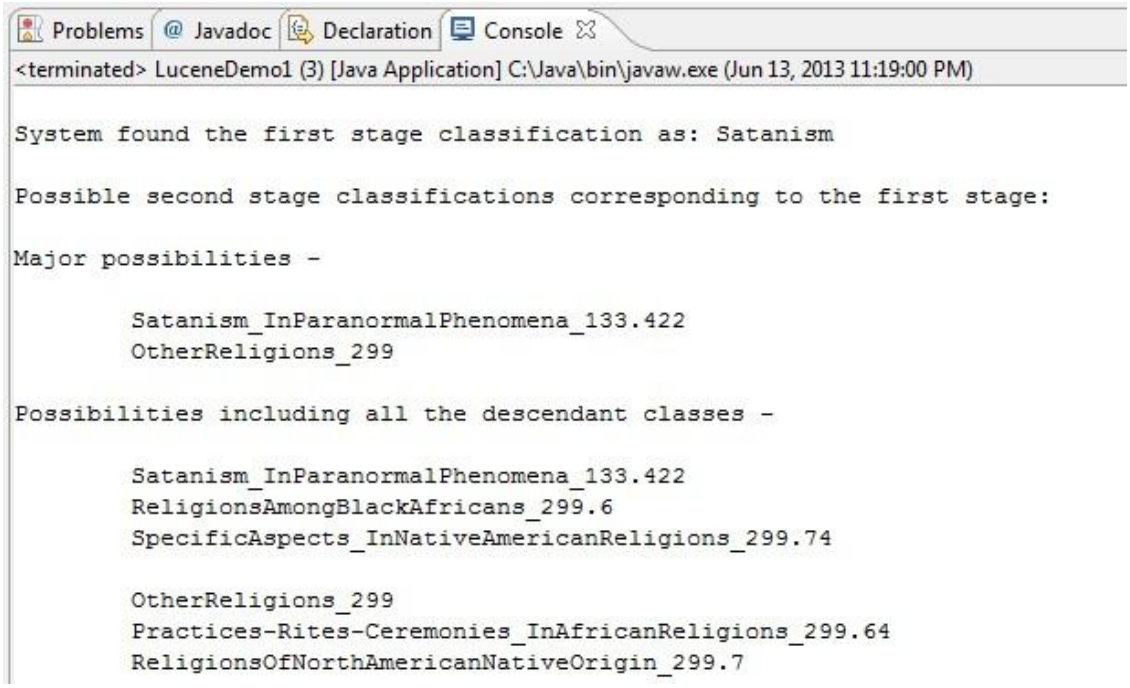

Figure 4: Classification results obtained after the second filtration

这 58 个文档再次应用到当前研究中获取分类结果。他们通过自动系统进行分类，系统仅给出了一个与输入文档最匹配的主题。结果通过两种方式获得，一是对比手工、半自动和全自动分类的准确性，二是评估文档模糊性和分类不准确性之间的关系，获得了手工、半自动和全自动三种分类方法的评估结果。

图 5 展示了新系统对“其他宗教”领域相同文档的分类结果。

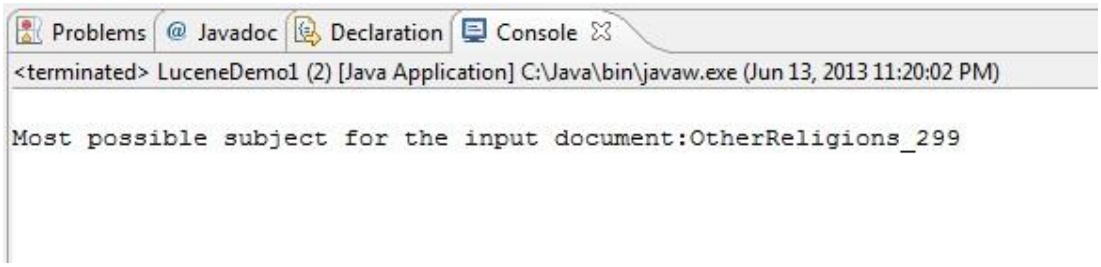

Figure 5: Classification results obtained by the fully-automatic system

4.1 分类准确性方面的结论

表 1 展示了手工、半自动、全自动三种分类方法的分类准确性对比结果。第一列给出了每一个主题中用于测试分类准确性的文档数量。第二列、第三列和第四列展示了每种分类法获得的正确分类的数量。

Table 1: Correct manual, hybrid and automatic classifications

Subject	Total Documents	Correct Manual Classifications	Correct Hybrid Classifications	Correct Fully Automatic Classifications

Apparitions	1	0	1	1
Aries	1	1	1	1
Attributes-Faculties	1	0	0	0
Axiology	1	1	1	1
Causation	1	0	1	1
Cosmology	4	2	3	0
Epistemology	2	1	2	0
Evil Spirits	1	0	0	0
Feng Shui	1	0	0	1
Geomancy	1	1	1	1
Leo	1	1	1	1
Libra	1	1	1	1
Love	1	1	1	1
Mind	1	0	0	1
Ontology	4	3	3	4
Other Religion	1	1	1	1
Palmistry	2	2	2	2
Phrenology	2	2	2	2
Pisces	1	1	1	1
Poltergeists	3	3	2	2
Precognition	2	2	2	2
Psychic Phenomena	2	2	2	0
Psycho Kinesis	3	2	2	3
Reincarnation	3	2	2	2
Space	1	1	1	1
Specific Mediumistic Phenomena	3	2	0	1
Spells-Curses-Charms	4	3	3	0
Spiritualism	1	0	1	1
Taurus	2	2	2	2
Teleology	1	1	1	1

Telepathy	3	3	2	2
Time	2	2	2	2

如表 1 所示，有 5 个文档采用全自动分类系统的分类结果比分类人员在自动系统协助下的分类结果准确性高，这些文档属于风水、心灵、本体、心理避世运动和具体的灵媒现象领域。另一方面，有 10 个文档采用混合系统的分类结果比全自动分类系统的分类结果准确性高，这些文档属于宇宙论，认识论，心灵现象，法术-咒语-护身符领域。但是，在所使用的全部主题中，利用全自动系统比混合方法的准确度高出 3.125%。

图 6 展示了半自动分类和全自动分类准确性百分比对比情况。

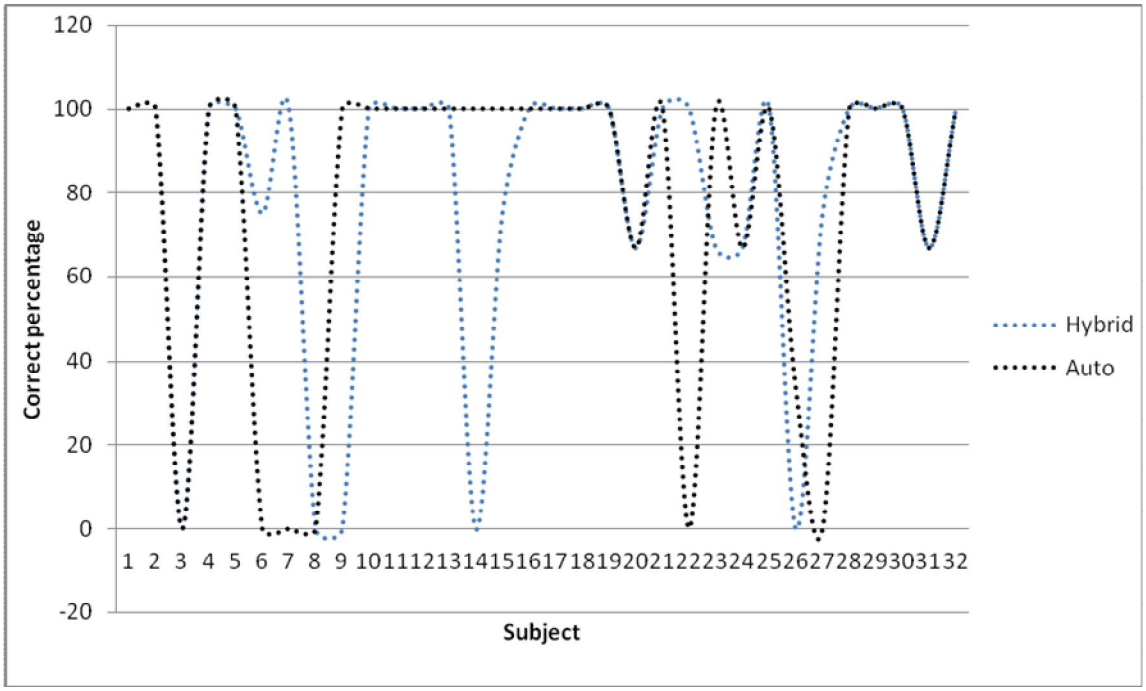

Figure 6: Percentages of correct semi-automatic and fully-automatic classifications

现在，我们注意到图 6 中两条分类曲线所显示的准确的平均分类数量差异并不明显。尽管在新系统中改进了准确性，但由于差异不显著导致图 6 中表现不明显。

4.2 文档模糊和误分类方面的结论

分析相同的文档集来检验文档模糊性和三种分类方法是否存在关系。另外，由分类人员确定文档是否存在模糊性。模糊的文档标记为 1，反之为 0。分类完成之后他们同样要被标记，分类不准确的文档标记为 1，分类准确的文档标记为 0。具体结果见表 2。

Table 2: Document vagueness and inaccurate classifications done by manual, hybrid and fully-automatic methods

Subject	Vagueness	Incorrect Manual Classification	Incorrect Hybrid Classification	Incorrect Fully Automatic Classification
Apparitions	1	1	0	0
Aries	0	0	0	0
Attributes Faculties	1	1	1	1
Axiology	0	0	0	0
Causation	1	1	0	0
Cosmology	1	1	0	1
	1	0	0	1
	1	1	1	1
	1	0	0	1
Epistemology	1	1	0	1
	0	0	0	1
Evil Spirits	1	1	1	1
Feng Shui	0	1	1	0
Geomancy	0	0	0	0
Leo	0	0	0	0
Libra	0	0	0	0
Love	1	0	0	0
Mind	1	1	1	0
Ontology	1	0	0	0
	1	1	1	0
	1	0	0	0
	1	0	0	0
Other Religion	1	0	0	0
Palmistry	0	0	0	0
	0	0	0	0
Phrenology	0	0	0	0
	0	0	0	0
Pisces	0	0	0	0
Poltergeists	0	0	0	0
	0	0	1	1
	0	0	0	0
Precognition	0	0	0	0
	0	0	0	0
Psychic Phenomena	1	0	0	1
	1	0	0	1
Psycho Kinesis	0	0	0	0
	0	0	0	0
	1	1	1	0
Reincarnation	0	0	1	1
	1	1	0	0
	0	0	0	0
Space	1	0	0	0
Specific Mediumistic Phenomena	1	0	1	0
	1	0	1	1
	1	1	1	1
Spells-Curses-Charms	0	0	0	1
	0	0	0	1
	0	0	0	1
	1	1	1	1
Spiritualism	1	1	0	0

Taurus	0 0	0 0	0 0	0 0
Teleology	0	0	0	0
Telepathy	0 0 0	0 0 0	1 0 0	1 0 0
Time	1 1	0 0	0 0	0 0

利用二元逻辑回归法对上述结果进行分析。比值比用于决定文档的模糊性和通过手工、半自动和全自动方法的分类结果非准确性之间是否存在关系。事实上，比值比用于测量影响的大小，描述两个二进制数据值之间的关联强度。

文档模糊性和手工分类非准确性之间的比值比是 29.00，文档模糊性和半自动分类非准确性之间的比值比是 3.61，文档模糊性和全自动分类非准确性之间的比值比是 2.46。因此，全自动方法分类中文档模糊性对误分类的影响最低。

5 结语

通过对比手工、半自动和全自动方法获得的分类结果确定分类效果。尽管半自动分类系统在文档数量上比全自动分类方法的准确性更高，但全自动分类系统在主题上的分类准确性比半自动分类高出 3.125%。原因可能是新系统中对某些学科的误判断，比如宇宙学和法术-咒语-护身符。因此，当属于这些学科的输入文档数量越多时，误分类的文档数量就越多。除了这一特殊情况外，新分类系统在分类准确性的其他所有方面都要好于之前的方法。此外，本研究表明，文档模糊性对分类准确性的影响在半自动分类方法中高于全自动分类方法。

参考文献

Abbas, M., Smaïli, K., & Berkani, D. (2010). Efficiency of TR-Classifier versus TFIDF. *2010 First International Conference on Integrated Intelligent Computing*. doi: 10.1109/ICIIC.2010.60

Best, B. J., Nathan, G., & Lebiere, C. (2010). Extracting the Ontological Structure of OpenCyc for Reuse and Portability of Cognitive Models. *Proceedings of the 19th Conference on Behavior Representation in Modeling & Simulation (BRiMS 2010)*. Retrieved from http://www.adcogsys.com/pubs/Brims2010-best-gerhart-lebiere-opencyc.pdf

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*. Retrieved from http://courses.media.mit.edu/2002fall/mas962/MAS962/miller.pdf

Morato, J., Marzal, M. A., Llorens, J., & Moreiro, J. (2004). WordNet Applications. *Proceedings of the 2nd International Conference on Global WordNet*. Retrieved from http://www.fi.muni.cz/gwc2004/proc/105.pdf

Prabowo, R., Jackson, M., Burden, P., & Knoell, H. D. (2002). Ontology-Based Automatic Classification for the Web Pages: Design, Implementation and Evaluation. *Proceedings of the 3rd International Conference on Web Information Systems Engineering*. Retrieved from http://portal.acm.org/citation.cfm?id=674083

Prevot, L., Borgo, S., & Oltramari, A. (2005). Interfacing Ontologies and Lexical Resources. *Proceedings of OntoLex 2005*. Retrieved from http://www.loa-cnr.it/Papers/%5B22%5DprevotBorgoOltramari-3.pdf

Song, M. H., Lim, S. Y., Kang, D. J., & Lee, S. J. (2005). Automatic Classification of Web Pages based on the Concept of Domain Ontology. *Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05)*, 645-651. doi: 10.1109/APSEC.2005.46

Tenenboim, L., Shapira, B., & Shoval, P. (2008). Ontology-based Classification of News in an Electronic Newspaper. *Proceedings of the International Conference on Intelligent Information and Engineering Systems*. Retrieved from http://www.foibg.com/ibs_isc/ibs-02/IBS-02-p12.pdf

Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing Affective Lexical Resources. *PsychNology Journal*, 2(1), 61-83. Retrieved from http://www.psychnology.org/File/PSYCHNOLOGY_JOURNAL_2_1_VALITUTTI.pdf

Wijewickrema, P. K. C. M. & Gamage, R. C. G. (2012). Automatic Document Classification Using a Domain Ontology, *Proceedings of the 09th National Conference on Library and Information Science (NACLIS 2012)*, ISSN 978-955-9075-17-2, 85-107.

Wijewickrema P. K. C. M. & Gamage, R. C. G. (2012). An enhanced text classifier for automatic document classification, *Journal of the University Librarians' Association of Sri Lanka, 16* (2), ISSN 1391-4081, 138-159.