1. **译文**

**结合语义和统计信息进行Web敏感文本过滤**

**摘要**

网络敏感信息是指网络上含有色情内容的文本、图片等形式的信息。如何过滤这些有害信息引起了研究人员的兴趣。为了保证网络内容的安全，各国政府也对这一问题的研究给予了极大的支持。本文首先简要回顾了网络敏感信息过滤的最新发展，然后分析了敏感文本的统计特征和语义特征，并用一个类似于CNN的词网来表示。最后，提出了一种语义与统计相结合的网络敏感文本过滤方法。实验结果证明了该方法的有效性。

1. **简介**

互联网为人们获取和交换信息提供了便利。然而，它也给我们带来了有害的内容，如色情、暴力和其他非法信息。这些有害的内容自然会对整个社会产生严重的影响，尤其是年轻人。因此，敏感信息的过滤具有重要的意义，是近年来研究的热点之一。

文献中已经有大量的过滤方法，大致可分为以下三类[1][2]。

1. **PICS** （因特网内容选择平台）本质上是一组用于对web站点进行评级的内容评级系统的规范。通常有两种方法来评价web页面。一种是自我评分，另一种是第三方评分，其区别在于评分结果是否由网络发布者给出。过滤系统可以通过检查网站的评级信息进行操作。
2. **黑名单和白名单**。由预先手动或自动生成网站列表，形成黑名单和白名单。黑名单记录禁止访问的网站的url。白名单记录允许访问的web站点的url。对于给定的新web页面，是否允许访问取决于所请求的URL与黑名单或白名单的匹配结果。
3. **关键字过滤**。这种方法的思想是，敏感文本总是包含一些特定的单词或短语，而它们通常不会出现在正常文本中。由这些特定的词或短语組成的单词列表，通常被用来构建关键字过滤方法，即计算语句中包含的词库匹配单词数，如果一个web页面匹配的单词数超过一个预定义的阈值，则不允许被浏览。

上述方法在敏感信息过滤方面都有各自的优点，但缺点也很明显。由于PICS不是强制性的标签系统，所以评级信息并不总是可用的。很难保持URL列表的完整和最新；因此，黑名单和白名单的方法不能有效地处理敏感页面。在基于关键字的过滤中，很多普通文本在单词列表中也包含一些特定的单词。因此，这种方法将不可避免地导致过度阻塞。

目前有许多商业网络过滤系统可用。2001年，欧盟委员会发起的网络保护项目[3]挑选了50个商业网络过滤系统并对其性能进行了评估。由于这些系统大多使用上述一种或多种传统方法，因此它们显然不能在实际应用中提供令人满意的结果。

为了更准确地过滤web上的敏感信息，近年来，研究人员将重点放在了智能内容识别方面，提出了多种检测成人图像[4]的算法。然而，他们只能在一定程度上识别某些类型的成人图像。其他一些研究人员更关注敏感文本的过滤[1][2]。Lee[1]等人在传统的基于关键字的过滤方法的基础上，通过计算文本中出现的关键字的数量得到一个特征向量，然后将该向量作为输入到KSOM神经网络中进行文本分类。尽管论文的结果表明这种方法是有效的，但当输入的文本是关于性健康和其他相关主题时，通常会给出错误的结果。Du[2]等人使用文本分类来过滤web上的敏感文本。在一个仅从Yahoo的成人类别中收集成人文本的测试数据集上，他们的方法获得了很高的准确性。事实上，情色故事和文本的风格并不相同，所以这种方法在现实世界中并不适用。

综上所述，在这方面还有三个主要问题没有很好地解决，即

**过度屏蔽问题**：如何区分敏感文本与性健康、文化等相关主题文本是一个具有挑战性的问题，许多方法都不能有效地解决它。

**拼写错误问题**：如果特定的单词有意或无意地拼写错误，许多方法可能无法正常工作。

**单词列表问题**：如何构造一个足够实用的单词列表是许多基于关键字的过滤方法的关键问题。然而，到目前为止，还没有人关注这个问题。

在这篇论文中，我们将专门把对敏感信息检测有用的单词分为三类。结合文本的语义和统计信息，得到了一种更有效的文本特征，用于敏感信息的过滤。本文的其余部分组织如下。第2节简要介绍了细胞神经网络(CNN)，分析了敏感文本的语义特征，并在第3节设计了一个类似于CNN的词网进行特征表示。第4节简单总结了该算法的主要步骤。实验结果在第5节给出并讨论，第6节给出最后的结论。

1. **细胞神经网络**

CNN是一个定义在离散n维空间[5]中的大规模并行计算范例，其中每个单元是一个多输入单输出处理器。细胞之间的连接构成了网络。CNN与其他神经网络的主要区别在于，只允许相邻细胞之间的连接，这使得我们可以通过局部交换和处理信息来获得全局的处理。图1显示了cnn的一个示例。

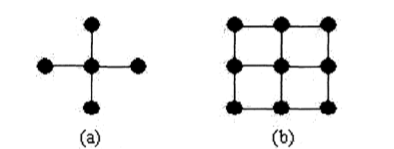


图1 (a) A细胞及其相邻细胞 (b) 3\*3大小的二维CNN

一个CNN动态系统可以在连续时间(CT-CNN)或离散时间(DT-CNN)[5]下工作。CNN中的每个单元格都有一个内部状态变量。细胞的内部状态，它的输出和来自相邻单元格的输出，三个部分，决定了它的最终输出。离散时间情况下的数学描述如下：

 (1)

x(t)是t时刻细胞的内部状态，y(t)是输出，u(t)是来自相邻单元格的外部输入，I(t)是一个叫做偏差的局部值。另外，f1和f2分别是两个参数函数。

CNN的理论在信号和图像处理[6][7]等领域得到了广泛的应用。在本研究中，我们将利用CNN的主要思想来构建一个类似于CNN的词网，来说明输入文本的语义特征。

1. **敏感文本的特征分析和表示**
2. **统计特征分析**

文本分类是将新文本分配到预定义的类别中。第一步也是最重要的一步，是将文本转换成合适的特征表示。定义文本特征的方法有很多，常用的方法是使用文本中出现的单词的统计数据。向量空间模型(Vector Space Model, VSM)可能是文本分类中最值得注意的模型，其中[8]，文档通常由单词向量表示。设A表示文本的特征，则

 (2)

其中，ai是单词i的权重，N是单词的数量。这里的关键步骤是如何定义单词的权重。Kerstin等人在他们的论文[8]中描述了6种不同的权重策略。fi是文本中单词i出现的频率。一种简单的方法是使用单词的频率作为其相关的权重，即

 (3)

敏感文本过滤的任务是确定输入文本是敏感的还是正常的，它可以看作是一个文本分类问题。现有的文本过滤方法大多基于这种思想，首先编译包含特定单词的单词列表。然后像(2)这样的向量被创建到文本中。显然，这样的向量是文本的统计特征。尽管这对文本分类是有用的，但是如果只使用这种统计方法，那么关于文本的所有语义信息就几乎不会被研究。

1. **语义特征分析**

一般来说，一些特定的词，如性和乳房，被认为是敏感文本的语义特征。事实上，许多与性有关但正常的文本也包含这些词。因此，它们可以提供错误线索来预测文本的类别。此外，如果出现拼写错误的问题，文本中的任何线索都不可能被正确收集。因此，如何从文本中提取正确的线索至关重要。

敏感文本中的词可能会给出与正常文本不同的语义。但是我们不知道输入文本在开始时是否敏感。其他信息，比如单词的上下文，会决定我们是否应该提取这些单词作为线索。我们知道，有些单词本身不包含任何敏感语义。但如果它们与其他一些词结合起来，就能提供敏感的线索。基于以上考虑，我们专门根据语义将对我们有用的词划分为以下三个类：

**明显的关键词(Obvious)**：这类词大约只出现在敏感的文本中。在统计学意义上，它们在普通文本中出现的概率接近于零。在语义上，它们代表情爱意义。

**隐藏的关键词(Hidden)**：这类词不包含色情含义。但由于某些原因，它们与敏感文本之间存在着混淆的关系。也就是说，它们出现在敏感文本中的概率很高，尽管它们也出现在正常文本中。

**合逻辑的关键词(Logical)**：这类关键字可以进一步分为两个子类。一种是乳房等多语义词，即它在正常文本中提供正常信息，在敏感文本中提供情色信息。另一种是，如果只有一些特定的词与它们相伴，这些词就被认为是情色关键词。

许多现有的方法只使用明显的关键字和逻辑关键字的第一个子类。实际上，隐藏关键字和逻辑关键字的第二个子类也可以极大地帮助对新文本进行分类。

人类大脑中有大量的单词，它们之间并不是相互孤立的。这些词之间的语义关系构成了一个巨大的网络，便于准确处理文本信息。例如，当我们在一篇文章中阅读一个单词时，我们可能会联想到其他语义相关的单词。此外，当我们读取一个单词时，该单词对应的节点接受它作为输入。然后根据节点之前的状态和相邻节点的状态改变节点的状态。这一机制启示我们，三类关键词及其语义关系能够合理地表征敏感文本的语义特征。在接下来的小节中，我们将解释它们之间的语义信息，并据此构建一个类似CNN的词网来描述它们的语义特征。图2显示了传统方法和我们的方法在关键字集或单词列表方面的主要区别。

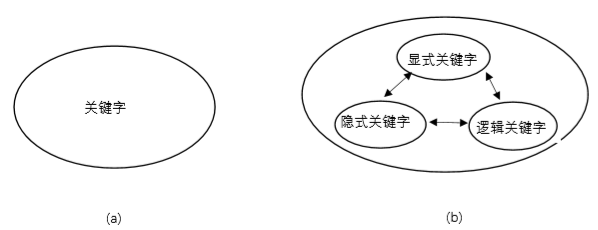


图2. 传统的关键字集和我们的关键字集

词汇之间存在语义信息。当我们看到“老师”这个词，“学生”这个词可能会在我们的脑海中无意识地出现。如果你在一段话中只看到三个单词“父母”，“学习”和“孩子”，你可以说这段话是关于父母在儿童教育中的角色。这意味着只有几个分开的单词才能给我们一个完整的意思。当你看到一个词“学生”，你可能不确定它是指中学生还是大学生。但是，如果你在下面的段落看到“学士”，你就可以说它最有可能是指大学生了。诸如此类的例子表明，单词之间的语义信息可以帮助我们获得更丰富的关于单词的线索。在本研究中，我们将尝试探索这些信息，以帮助我们从文本中提取正确的线索。

为了构造类似于CNN的单词网络，我们将一个单元定义为一个单词。用四个参数来描述一个单元格。它们分别是状态、外观位置集、数量和独立的激活数量。细胞有三种状态：如果不出现就处于睡眠状态，如果出现但未被激活就处于休眠状态，如果被激活就处于活动状态。单元格的输出等于单元格的内部状态，而单元格的内部状态由上述四个参数描述。我们不使用严格的函数来描述单元格的输入和输出，而是使用一些语义规则。一个单元格及其相邻单元格如图3所示。

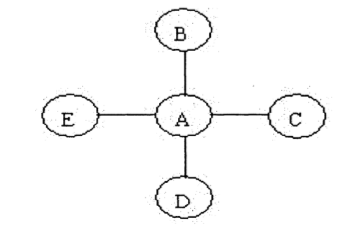


图3. 一个单词细胞元及其相邻的单词细胞

假设A是一个隐藏的关键字，它的初始状态是休眠。当有一个针对A的输入时，如果B、C、D、E的参数以及之前A的参数满足一定的规则时，A就会被激活。否则，节点A将变为休眠状态。这个过程可以帮助我们有效地提取敏感词。例如，在一篇关于性健康的文章中，虽然A也可能出现，但是参数不能满足一定的规则，那么A就不能被激活。如果有与A相似的单词，且参数符合规则，我们也可以激活A。只有那些被激活的词才被认为是有用的线索。如果A是一个明显的关键字或合逻辑的关键字，其处理方法与此类似。很明显，这个过程可以帮助解决拼写错误的问题。因为当我们看到拼写错误的单词时，我们可以返回上下文或其他语义信息来正确地理解它。

整个类似于CNN的词网是由上面描述的一些范例构建的。任何两个单元之间的链接都显示了它们的语义关系。规则指示在什么情况下可以激活细胞。不同的范式有不同的规则。三种词语之间的联系和规则可以有效地表征敏感语篇的语义特征。

1. **算法**

该方法的关键步骤是正确构建类CNN的词网。关键字是根据我们的关键字分类策略选择的。理想情况下，我们最好使用机器学习来自动获取细胞（单词）之间的规则（语义关系）。考虑到只需要过滤特定的文本，这里我们手动构造类似于CNN的单词网络。

1. **特征抽取**

所有细胞的初始状态都设置为睡眠状态。特征提取涉及的主要步骤总结如下。

**S1.** 从输入文本中获取一个单词，并逐渐将该单词与类似CNN的网络中的每个单元格匹配。找到一个与输入单词相似度最高的单元格。如果相似度评分超过预先设定的阈值，则调整该单元格外观位置集、数量的参数，然后转到S2;否则，转向SI。

**S2.** 从类CNN网络中获取细胞及其邻近细胞的参数。如果这些参数满足一定的规则，那么这个单元格就会被激活。它的激活号被添加到1，它的状态变为活动。否则，就会变成休眠。

**S3.** 改变状态，调整相邻单元格的状态，过程与S2相同。然后迭代调整整个网络。如果所有文字都已经被处理，转到S4；否则就转到S1。

**S4.** 收集每个细胞的激活数量并形成一个向量。

向量用于表示输入文本的语义特征和统计特征。

1. **训练和分类**

支持向量机（Support vector machine，SVM）是目前非常流行的分类技术。它将分类问题转化为线性布局问题。该算法在训练数据的不同类别之间找到一个超平面。一旦确定了超平面，我们就可以使用它对一个新的数据[2]进行分类。支持向量机同样适用于文本分类[9]。基于SVM在文本分类中的良好性能，我们选择SVM作为我们的分类器。

1. **实验评估**

为了评估该方法的效果，我们从互联网上收集了3162条中文文本，其中包括577条敏感文本、585条与性有关但正常的文本和2000条正常文本。普通文本包括艺术、商业、科学、计算机、新闻、购物、游戏娱乐、社会、健康和体育等10个子类。每个子类别包含200个web文本。健康分类主要来自以下网站:www.xyxy.net、health.21cn.com、www.fml20.com和www.medicch- ina.com。与其他网页过滤文本数据库相比，我们只收集与性相关的正常样本，如性健康、性文化和性教育。300个敏感文本，300个与性别相关的正常文本和1000个普通文本用作训练数据，其余用作测试数据。编制了109个指示性词汇表，包括29个明显关键词、33个隐藏关键词和47个逻辑关键词。构造了一些简单的规则来描述这三类关键字之间的关系。所有的关键字和简单的规则构成了类似于CNN的词网。

设计了三种不同的特征提取方案来测试该方法的有效性。第一种是传统的方案，只计算每个明显关键字的个数和一部分逻辑关键字。第二步是统计这三种关键字。第三种是通过类似CNN的词网来计算关键词。自由软件Libsvm-2.6[10]用于训练和预测我们的数据库。表1总结了实验结果，从表中数据和虚假的文本，我们可以得到一些有用的结论：

(1)将方案1与方案2进行比较，我们发现我们定义的三种关键字，并根据该定义构造关键字集，可以显著提高识别率。但由于关键字集包含许多逻辑关键字和隐藏关键字，因此将一些普通文本分类为敏感类是不可避免的。

(2)利用类CNN词网提取文本特征，得到最佳分类率。证明了类CNN词网能够较好地表征敏感文本的语义特征。

(3)真实敏感文本的文体和内容千变万化。因此，为了获得更高的分类率，我们需要扩大关键字集。

(4)由于我们只考虑单词之间的一些朴素语义关系，一些正常的相关文本也被预测为敏感文本。

表1 三种方案的分辨率

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 敏感文本 | 相关文本 | 正常文本 | 合计 |
| 方案1 | 93.14% | 95.08% | 100% | 97.88% |
| 方案2 | 96.38% | 97.54% | 99.80% | 98.78% |
| 方案3 | 97.83% | 98.24% | 100% | 99.29% |

Du等人使用文本类别过滤敏感网页[2]。虽然他们的训练和测试集是英文文本，而我们的是中文文本，但是我们将他们的方法与我们的方法进行了比较。因为它们和我们的方法都不是面向特定语言的。在他们的方法中，需要手动设置阈值t。表2给出了两种方法的结果，其中阻塞率为敏感文本集中正确分类文本的比例，而过阻塞率为非敏感文本集中错误分类文本的比例。Du的方法的结果直接取自他的论文。可以看出，虽然我们的文本数据集比他们的更具挑战性，但是我们的结果在整体上还是比他们的好。他们收集了测试数据仅来自雅虎的关于成人的分类，并不包含与性别相关的正常文本。

表2 三种方法的分辨率

|  |  |  |
| --- | --- | --- |
|  | 阻塞率 | 过阻塞率 |
| 我们的方法 | 97.83% | 0.39% |
| Du的方法（t=0.18） | 97.41% | 0.48% |
| Du的方法（t=0.10） | 99.35% | 4.09% |

实验结果也表明，该方法能有效地解决语言的过阻塞问题和单词列表问题。我们对关键字的分类可以指导我们如何选择关键字来更有效地构造单词表(或者关键字集)。我们的类似CNN的词网可以帮助从文本中提取正确的线索，避免阻塞正常的文本。我们还没有做关于拼写错误问题的实验。对于英语单词来说，计算两个相似单词之间的相似性是很容易的。例如，单词“University”可以拼写为“Uinervtisy”。 但是，由于汉语单词在发音或形状上可能是相似的，所以处理起来相对比较困难。但可以相信，如果有一种简单的方法来计算汉语单词之间的相似性，那么该方法可以解决拼写错误的问题。

1. **结论**

本文定义了三种关键字，构建了一个类似于CNN的词网来提取和表示文本的语义特征和统计特征。本研究试图运用语义学来解决这一领域中尚未解决的三个问题。实验结果表明，该方法具有良好的应用前景。今后的工作将集中在

(1)扩大关键字集合，设计更准确的词与词之间的语义规则，构建更好的类CNN词网

(2)寻找一种可行的方法来计算两个中文单词之间的相似性。

**致谢**

我们非常感谢Jeffrey L. Wang(伦敦帝国理工学院计算机系)用英语提出的修改意见和宝贵建议。同时，我们也要感谢中国科学院模式识别国家实验室的陈明亮和陈周耀在网页采集方面的辛勤工作。

**参考文献**

1. P.Y. Lee, S.C. Hui and A. Fong, "Neural Networks for Web Content Filtering", IntelligentSystems, 17(5): 48-57, 2002
2. R. Du, R. Safavi-Naini and W. Susilo, "Web Filtering Using Text Classification", Proc. the 11th IEEE Intl. Conf on Netvork, pp. 325 330, 2003.
3. NetProtect Research Group, "Report in Filtering Techniques and Approaches NETPROTECT: WP2: 2.3 VI.023", TechnicalReport, Oct 2001
4. D. Forsyth and M. Fleck, "Automatic Detection of Human Nudes", InternationalJournalofComputer Vision, 32 (1): 63-77, 1999.
5. <http://www.ce.unipr.it/pardis/CNN/#InterPoint>
6. A. Lukianiuk, "Capacity of Cellular Neural Networks as Associative Memories ", Proc. the Fourth IEEE Int. Workshop on Cellular Neural NetworksandTheirApplications, pp. 3740, 1996.
7. M. G. Milanova, A. C. Campilho, and M. V. Correia, "Cellular Neural Networks for Segmentation of Image Sequence", Proc. the 11th Portuguese Conference on Pattern Recognition, pp. 49-54, 2000.
8. K. Aas and L. Eikvil, "Text Categorization: a Survey", TechnicalReport -941, Norwegian ComputingCenter, 1999.
9. T. Joachims, "Text Categorization with SupportVectorMachines: Learningwith Many Relevant Features", Proc. the 10th European Conference on Machine Learning, pp.137-142, 1998.
10. http://www sie.ntu.edu.tw/-cjlin/papers/guide/guide.pdf
11. **原文**

**Web Sensitive Text Filtering by Combing Semantics and Statistics**

**Abstract**

Web sensitive information is defined as texts, pictures and other forms of information which contain erotic content on web. How to filter this harmful information attracts researchers' interests. In order to keep web content safe, governments have also given great support on the research on this problem. This paper first briefly review recent developments in web sensitive information filtering then the statistic and semantic features of sensitive texts are analyzed and represented by a CNN-like word net. Finally a novel method which combines semantics and statistics is proposed to filter sensitive text on web. Experimental results have demonstrated the proposed method's promising performance.

**I. Introduction**

Internet has facilitated ones to obtain and exchange information. However, it also brings us harmful contents such as pornography, violence and other illegal messages. These harmful contents naturally have serious influence on the whole society, especially young people. So sensitive information filtering is ofgreat importance, and has been one ofmost active researchtopics recently.

There have been a large number offiltering methods in the literature, which can be roughly classified into three major classes as follows [1][2].

PICS (Platform for Internet Content Selection) is in essence a set ofspecifications for content-rating systems that can rate web sites. There are usually two measures to rate the web pages. One is Self-rating and the other is the Third-part rating, which are distinguished by whether the rating results are given by web publishers or not. The filtering systems can operate through checking this rating information ofweb sites. Blacklists and Whitelists.

Blacklists and Whitelists are lists ofweb sites which compiled manually or automatically beforehand. Blacklists record URLs of web sites which are forbidden to access. Whitelists record URLs of web sites which are allowed to access. For a given a new web page, whether it is allowed to access or not depends on the matching result ofthe requested URL with the Blacklists or Whitelists.

Keyword-based Filtering. The idea of this approach is that sensitive texts always contain some specific words or phrases while they do not usually appear in normal texts. A word list that is composed ofthese specific words or phrases is usually needed to be constructed for keyword-based filtering methods, which count he number ofwords contained in the wordlist by matching the word list and a web page and do not allowed to browse when the number exceeds a predefined threshold.

Each kind ofmethod mentioned above does have its own advantages in sensitive information filtering, but their drawbacks are also obvious. The PICS is not a compulsory labeling system, so the rating information is not always available. It is very difficult to keep the URL lists complete and up to date; the approach ofBlacklists and Whitelists thus cannot deal with the sensitive pages effectively. As to the Keyword-based Filtering, many normal texts also contain some specific words in the word list. Therefore, this approach will leadto overblocking inevitably.

There are many commercial Web-filtering systems available currently. In 2001, The NetProtect project [3] launched by European Commission selected fifty commercial web-filtering systems and evaluated their performance. Because most of these systems use one or more traditional approaches above, it is clear that they cannot provide satisfactory results in real applications.

In order to filter the sensitive information on web more accurately, researchers have recently focused on research on the intelligent content recognition. Various algorithms are proposed to detect adult images [4]. However, they can only recognize certain kinds ofadult images to some extent. Some other researchers have paid more attention to sensitive text filtering [1][2]. Based on the traditional approach of Keyword-based Filtering, Lee et al. [1] counted the number ofkey words appearing in the text to obtain a feature vector, and then used the vector as the input into a KSOM neural network for text classification. Although the results in the paperhave shownthatthis method is efficient, itusually gives wrong results when the input text is about sexual healthy and other related topics. Du et al. [2] used text classification to filter sensitive texts on web. On a test data set in which adult texts were collected only from the adult category of Yahoo, their method achieved a high accuracy. In fact, the styles of erotic stories and texts are not in common, so this approach cannotworkwell inthe real world.

In summary, there are three major problems which are not well solved in this area, i.e.

Overblocking problem: How to distinguish sensitive texts from related topic texts such as sex-related health and culture is a challenging problem which many methods can not solve efficiently.

Mis-spelled problem: Many approaches probably cannot work normally if the specific words are misspelled intentionally orunintentionally.

Wordlist problem: How to construct a sufficient and practical wordlist is a key problem for many keyword-based filtering approaches. However, so far, nobody has focused on this problem.

In this paper, we will specifically divide words which are useful for sensitive information detection into three classes. By combining semantics and statistics of texts, a more efficient text feature is obtained for the purpose of sensitive information filtering. The remainder ofthis paper is organized as follows. Section 2 briefly introduces the Cellular Neural Network (CNN), and semantic features of sensitive texts are analyzed and a CNN-like word net is designed for feature representation in Section 3. Section 4 simply summaries major steps of the proposed algorithm. Experimental results are given and discussed in Section 5, prior to conclusions in Section 6.

**II. Cellular Neural Network**

CNN is a massive parallel computing paradigms defined in discrete N-dimensional spaces [5], in which each cell is a multiple input-single output processor. Cells and connections among them form the network. The main difference between the CNN and other neural networks is that connections are only allowed between adjacent cells, which allows obtaining global processing by exchanging and processing information in a local manner. Figure 1 shows an example ofCNN.

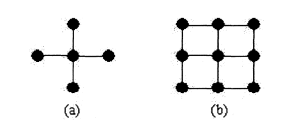


Fig.1. (a) A cell and its adjacent cells and (b) a two-dimensional CNN of 3\*3 size.

A CNN dynamical system can operate in continuous (CTCNN) or discrete time (DT-CNN) [5]. Each cell in a CNN is characterized by an internal state variable. Three parts, namely its internal state, its output and outputs from its adjacent cells, decide its final output. A mathematical description in a discrete time case is as follows:



where x(t) is the internal state of a cell in time t. y(t) is the output. u(t) is external input from adjacent cells and I(t) is a local value called bias. In addition, Jf and f2 are twoparametric functions respectively.

The theory ofthe CNN has been widely applied in many areas such as signal and image processing [6][7]. In this study, we will use the main idea ofthe CNN to construct a CNN-like word net to illustrate semantic features ofthe inputtexts.

**III. Feature Analysis And Representation For Sensitive Text**

1. Statistic feature analysis

Text categorization is to assign a new text into the predefined categories. The first and also predominate step is to transform texts into a suitable feature representation. There are many methods to define text feature. The common method is using statistical data of words appearing in text. Vector Space Model (VSM) may be the most notable model in text categorization, in which [8], documents are generally represented by vector ofwords. Let A denote the feature ofa text,

A = (al, a2. . . ai, an ) (2)

where ai is the weight of word i and N is the number of words we will count. The key step here is how to define the words' weights. Kerstin et al. [8] described 6 different weighting strategies in their paper. Let fJ be the frequency of word i in the text. A simple approach is to use the frequency ofword as its associated weight, i.e.

ai = fi (3)

The task of sensitive text filtering is to determine an input text sensitive or normal, it may be considered as a text categorization problem. Most of the existing texts filtering approaches are based on this idea, in which a wordlist which contains some specific words is firstly compiled. Then a vector like (2) is created to the text. Obviously, such vector is the statistical feature ofa text. Although it is useful to classify the text, almost all ofthe semantic information about the text is notyet explored ifonly usingthis statistical method.

1. Semanticfeature analysis

Generally, some specific words such as sex and breast are considered as the semantic features ofa sensitive text. In fact, many sex-related but normal texts also contain these words. So they may provide error clues to predict the category of a text. In addition, if the Miss-spelled problem occurs, any clues from the text will impossibly be collected correctly. So howto extract right clues from a text willbe critical.

Words in sensitive texts may give different semantics from they give in normal texts. But we don't know whether the input text is sensitive or not in the beginning. Other information such as context ofwords will be needed to decide whether we should extract these words as clues or not. As we know, some words do not contain any sensitive semantics by themselves. But ifthey are combined with some other words, they can provide sensitive clues. Based on the above consideration, here we specifically divide words which useful for us to filter into three classes according to semantics as follows:

**Obvious Keywords**: This class ofwords approximately only appears in sensitive texts. In a statistical sense, the probabilities oftheir appearances in normal texts are close to zero. In a semantic sense, they represent erotic meaning.

**Hidden Keywords**: This class of words does not contain erotic meanings. But for some reasons, there are confused relations between them and sensitive texts. That is to say, the probabilities they appear in sensitive texts are high though they also appear in normal texts.

**Logical Keywords**: This class of keywords can be further divided into two subclasses. One is multi-semantic word such as breast, which provides normal information in normal texts and erotic information in sensitive texts. The other is that ifonly some specific words are companied with them, these words are considered as erotic keywords.

Many existing approaches only use obvious keywords and the first subclass of logical keywords. In fact, hidden keywords and the second subclass of logical keywords also can greatlyhelp classify anew text.

There are giant numbers ofwords in human brains and they are not isolated each other. These words form a huge net by semantic relations among them, which will facilitate to process text information accurately. For example, when we read a word in an article, we may associate with other semantic related words. In addition, when we read a word, the node corresponding to this word accepts it as an input. Then the node's state is changed according to its previous state and the states ofits adjacent nodes. This mechanism enlightens us that the three classes ofkeywords and their semantic relations can represent the semantic features of the sensitive texts reasonably. In the next subsection, we explain the semantic information among them and accordingly construct a CNNlike word net to describe the semantic features. Figure 2 shows the main difference in keyword set or wordlist between the traditional and our approaches.

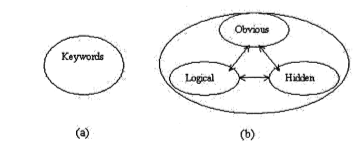


Fig.2. Traditional keyword setand ourkeyword set.

1. CNN-likewordnet

There exists semantic information among words. When we see the word "teacher", the word "student" may appear in our minds unconsciously. Provided that you only see three words "parents", "study" and "children" in aparagraph, you may say this paragraph is about parents' role in children's education. It means that only several separated words can give us an integrative meaning. When you see a word "students", you may be not sure whether it means middle school students or university students. But, if you see "bachelor" in the following paragraph, you can say it most possibly means university students. Examples like these have shown that semantic information among words can help obtain more informative clues about words. In this study, we will try to explore this information to help us extract right clues from a text.

To construct a CNN-like word net, we define a cell as a word. And four parameters are used to describe a cell. They are state, position set for appearances, number and activated numberrespectively. A cell hasthree kinds ofstates: sleep ifit doesn't appear, fallow ifithas appeared but not been activated and activity if it is activated. The output of a cell is equal to the internal state and the internal state ofa cell is described by the four parameters above. Instead ofusing strict function to describe the input and output of a cell, we just use some semantic rules. A cell and its adjacent cells are shown in Figure 3.

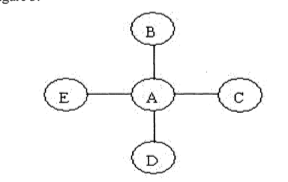


Fig.3. A word cell and its adjacent cells.

Provided that A is a hidden keyword and its initial state is sleep. When there is an input forA, ifthe parameters ofB, C, D and E and previous parameters ofA meet a certain rule, then A will be activated. Otherwise the node A turns to the fallow state. This process can help us extract sensitive-prone words efficiently. For example, in a text about sex healthy, although A may also appear, but the parameters can't meet a certain rule, thenA can'tbe activated. Ifthere is aword similar toA, andthe parameters meet the rule, we can also activate A. Onlythose activatedwords are considered as useful clues. IfA is an obvious or logical keyword, the disposal is similarto this. It is obvious that this process can help solve the Miss-spelled problem. Because when we see a miss-spelled word, we can return to its context or other semantic information to understand itcorrectly.

The whole CNN-like word net is constructed by a number of paradigms described above. Link between any two cells shows their semantic relations. Rules indicate in what case a cell can be activated. Differentparadigms have different rules. Links and rules among the three kinds of words can effectively representthe semantic features ofsensitive texts.

**IV. The Algorithm**

The key step ofthe proposed approach is constructing the CNN-like wordnetproperly. Keywords are selected according to our keyword classification strategy. Ideally, we had better use machine-learning to automatically attain rules (semantic relations) among cells (words). Considering that only specific texts needto be filtered, here we construct the CNN-like word netmanually.

1. Feature extraction

The initial states of all cells are set to sleep. Major steps involved in feature extraction are summarized as follows.

**S1.** Obtain a word from the input text and match the word with each cell inthe CNN-like netgradually. Find a cell which has the highest similarity with the input word. Ifthe similarity score exceeds a predefined threshold, adjust the parameters except the state and activated number of this cell, and then turnto S2; else, turn to SI.

**S2.** Get the parameters of the cell and its adjacent cells fromthe CNN-like net. Ifthese parameters meet a certain rule, then this cell is activated. Its activated number is added to one and its state turns to activity. Otherwise, itturns to fallow.

**S3.** Ifits state is changed, adjust states ofthe adjacent cells by the same process as S2. Then adjust the whole net iteratively. Ifallwords ofthe texthave been processed, turn to S4; else turn to 51.

**S4.** Collect the activated number of each cell and forms a vector.

The vector is used to represent semantic and statistic features ofthe inputtext.

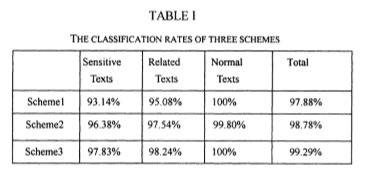
1. Trainingandclass.ification

Support vector machine (SVM) is a very popular classification technique now. It transforms classification to a lineal layout problem. The algorithm finds a hyper plane between different classes ofthe training data. Once the hyper plane is determined, we can use it to classify a new data [2]. SVM is also applied in text classification [9]. We choose SVM as our classifier for its well performance in text classification.

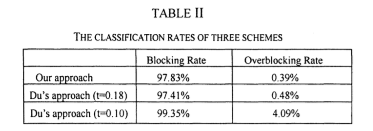
**V. Experiment**

To evaluate the proposed method's performance, 3162 Chinese texts have been collected from Internet, which include 577 sensitive texts, 585 sex-related but normal texts and 2000 normal texts. The normal texts consist of 10 subcategories, namely Arts, Business, Science, Computer, News, Shopping, Game & Recreation, Society, Health and Sports. Each subcategory contains 200 web texts. The Health subcategory was mainly collected from the following web sites: www.xyxy.net, health.21cn.com, www.finl2O.com and www.medicch- ina.com. Compared to other text databases for web filtering, only we collected sex-related normal samples such as sexualhealth, sexualculture and sexual education. 300 sensitive texts, 300 sex-related normal texts and 1000 normal texts are used as training data, andthe remaining serves as test data. A list of 109 indicative terms comprising 29 obvious keywords, 33 hidden keywords and 47 logical keywords has been compiled. Some simple rules are constructed to describe the relationships among these three kinds of keywords. All keywords and simple rules form the CNN-like word net.

Three different feature extraction schemes are designed to test the effectiveness of our approach. The first is the traditional scheme, in which only the number ofeach obvious keyword and a part of logical keywords is counted. The second is to count all ofthe three kinds ofkeywords. The third is to count keywords through the CNN-like word net. The free software Libsvm-2.6 [10] is used to train and predict on our database. The experimental results are summarized in Table 1, from which and false recognized texts, we can get several useful conclusions: (1) Comparing scheme 1 with scheme 2, we see that our definition three kinds of keywords and constructing the keyword set according to this definition can improve the recognition rates noticeably. But because the keyword set contains many logical keywords and hidden keywords, it is inevitable to classify some normal texts into sensitive category. (2) Using the CNN-like word net to extract features oftexts, we get the best classification rates. It proves that CNN-like word net can represent the semantic feature of sensitive texts properly. (3) The real sensitive texts have protean styles and contents. So in order to get a higher classification rate, we need to enlarge the keyword set. (4) Because we only consider some naive semantic relations among words, some normal related texts are also predicted as sensitive texts.



Du et al. [2] applied text category to filter sensitive web pages. Although their training and test sets are English texts and ours are Chinese, we make comparison between their approach and ours. Because both their and our methods are not specific language-oriented. In their approach, it needs to set the threshold t manually. Table 2 shows the results ofthe two approaches, where Blocking Rate is the fraction of the correct classified texts in the sensitive text set, while Overblocking Rate is the fraction ofthe false classified texts in the non-sensitive text set. The results of Du's method are taken directly from his paper. It can be seen that our result is better than theirs in an overall manner, even though our text data set is more challenging than theirs. They collected test data about adult only from the adult category of Yahoo and didn't contain sex-related normal texts.



Experimental results also show that that our approach can solve the Overblocking problem and Wordlist problem efficiently.Our classification ofkeywords can guide us how to select keywords to construct wordlist (or keyword set) more efficiently. Our CNN-like word net can help extract right clues from text and avoid blocking normal texts. We have not yet done experiments about the Mss-spelled problem. To English words, it is easy to calculate the similarity between two similar words. For example, the words 'University' may be spelled as 'Uinervtisy'. But it is relatively more difficult to handle the Chinese words because they may be similar in pronunciation or shape. But it is believable that, ifthere is a simple method to calculate the similarity between the Chinese words, the proposed approach can solve the Miss-spelled problem.

**VI. Conclusions**

This paper have defined three kinds of keywords and constructed a CNN-like word net to extract and represent semantic and statistic features of texts. This study is an attempt to use semantics to solve the three unsolved problems in this area. Experimental results have shown that the proposed approach is very promising. Future work will focus on (I) Enlarging our keyword set and designing more accurate semantic rules among words so as to construct a better CNN-like word net and (2) Finding a feasible way to calculatethe similaritybetweentwo Chinese words.

**Acknowledgement**

We are very appreciative to Jeffrey L. Wang (Dept.of Computing, Imperial College London) for his revision and valuable suggestions in English. We also thank MingliangZhu and Zhouyao Chen (National Laboratory of Pattern Recognition, CAS) fortheirhard work in web page collection.

**References**

1. P.Y. Lee, S.C. Hui and A. Fong, "Neural Networks for Web Content Filtering", IntelligentSystems, 17(5): 48-57, 2002
2. R. Du, R. Safavi-Naini and W. Susilo, "Web Filtering Using Text Classification", Proc. the 11th IEEE Intl. Conf on Netvork, pp. 325 330, 2003.
3. NetProtect Research Group, "Report in Filtering Techniques and Approaches NETPROTECT: WP2: 2.3 VI.023", TechnicalReport, Oct 2001
4. D. Forsyth and M. Fleck, "Automatic Detection of Human Nudes", InternationalJournalofComputer Vision, 32 (1): 63-77, 1999.
5. <http://www.ce.unipr.it/pardis/CNN/#InterPoint>
6. A. Lukianiuk, "Capacity of Cellular Neural Networks as Associative Memories ", Proc. the Fourth IEEE Int. Workshop on Cellular Neural NetworksandTheirApplications, pp. 3740, 1996.
7. M. G. Milanova, A. C. Campilho, and M. V. Correia, "Cellular Neural Networks for Segmentation of Image Sequence", Proc. the 11th Portuguese Conference on Pattern Recognition, pp. 49-54, 2000.
8. K. Aas and L. Eikvil, "Text Categorization: a Survey", TechnicalReport -941, Norwegian ComputingCenter, 1999.
9. T. Joachims, "Text Categorization with SupportVectorMachines: Learningwith Many Relevant Features", Proc. the 10th European Conference on Machine Learning, pp.137-142, 1998.
10. http://www sie.ntu.edu.tw/-cjlin/papers/guide/guide.pdf