**Computer Science and Application** 计算机科学与应用, 2024, 14(7), 35-41
Published Online July 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.147161

Hans汉斯

# 基于Transformer的服务推荐方法研究

**程芳颖，俞　婷，林帅伽，王雅琦，余振洋，陈鑫磊**

嘉兴南湖学院信息工程学院，浙江　嘉兴

收稿日期：2024年6月3日；录用日期：2024年7月5日；发布日期：2024年7月12日

## 摘　要

**近年来，越来越多的开发者使用Web服务来进行应用开发，但是如何选择合适的Web对于开发者来说存在着一定的难度。开发者对于各类服务不熟悉，无法精确对服务特征进行描述。因此，本文提出了一种基于Transformer的服务推荐方法(SRT)，首先，我们使用Transformer来对开发者提出的开发需求进行文本特征提取，接着我们使用深度神经网络来进一步挖掘应用和服务的潜在关系，进而进行服务推荐。在ProgrammableWeb上收集的真实数据集上进行的大量实验表明了我们所提出的SRT方法的有效性。**

## 关键词

**Transformer，服务推荐，文本特征**

# Service Recommendation Method Based on Transformer

**Fangying Cheng, Ting Yu, Shuaijia Lin, Yaqi Wang, Zhenyang Yu, Xinlei Chen**

School of Information Engineering, Jiaxing Nanhu University, Jiaxing Zhejiang

Received: Jun. 3rd, 2024; accepted: Jul. 5th, 2024; published: Jul. 12th, 2024

## Abstract

**In recent years, more and more developers have been using Web services for application development. However, it can be challenging for developers to choose the right Web services. They may not be familiar with various services and find it difficult to accurately describe their features. Therefore, we propose a Service Recommendation method based on Transformer (SRT). Firstly, we employ Transformer to extract textual features from the development requirements that provided by developers. Then, we utilize deep neural networks to further explore the potential relationship between applications and services, enabling service recommendations. Extensive experiments conducted on a real dataset collected from ProgrammableWeb demonstrate the effectiveness of our proposed method.**

文章引用: 程芳颖, 俞婷, 林帅伽, 王雅琦, 余振洋, 陈鑫磊. 基于 Transformer 的服务推荐方法研究[J]. 计算机科学与应用, 2024, 14(7): 35-41. DOI: 10.12677/csa.2024.147161

**Keywords**

**Transformer, Service Recommendation, Textual Features**



## 1. 引言

随着云计算、大数据技术的迅猛发展，Web 服务的数量和类型显著增加。例如，截至 2021 年 4 月，最大的 Web 服务目录之一 ProgrammableWeb 拥有 23,000 多个 Web 服务和 7,000 多个应用。在同一时期，流行的网络 API 市场 RapidAPI 列出了 20,000 多个 Web 服务。用户在进行应用开发时，通常会集成现有的 Web 服务，这样子，开发人员不必从头编写相关代码，从而减轻了开发时间和成本。如图 1 所示，应用程序 nearplace 是免费的商店定位器和谷歌地图标记，包含了不同的 Web 服务，其中有 MetaLocator、WordPress.org、Google Maps、WooCommerce、Shopify Admin 和 Magento SOAP。

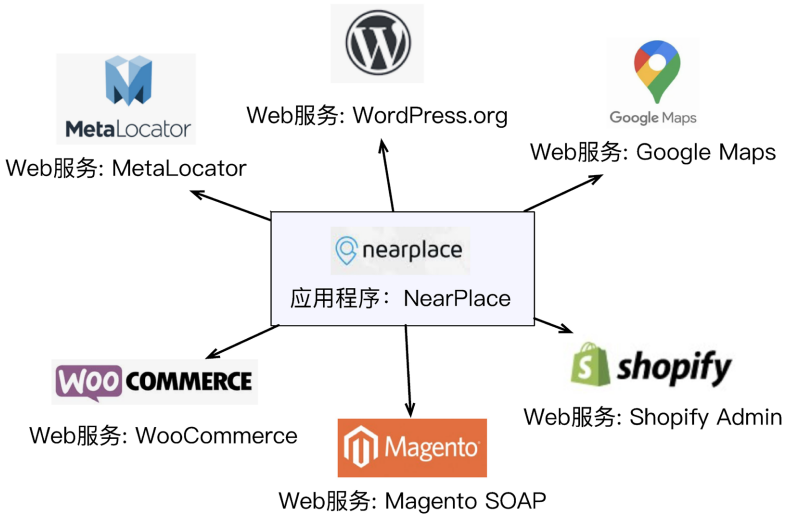


**Figure 1.** The example of an application
**图 1.** 应用程序示例图

通常，开发人员从 Web 服务库中浏览和选择相关服务，然后将它们集中到相关的应用程序中。这里的困难点在于，Web 服务库中存在着大量的候选 Web 服务，如何进行精准选择。因此，Web 服务的数量的不断增长对于 Web 服务推荐带来了一系列的挑战。

总结来看，Web 服务推荐存在着以下问题：

1) 通常，开发者输入开发需求来描述所要开发的应用，然而，用户的输入通常是随意的。让用户选择合适的词语来精准的表达开发需求是不太现实的，因为开发人员并不是 Web 服务领域的专家。因此，如何引入自然语言处理技术来尽可能准确的提取开发需求的特征是遇到的第一个问题。

2) 当前的 Web 服务库中存在着大量的 Web 服务和应用的历史交互，如何利用现有的交互历史记录来挖掘待开发应用和服务之间的关联关系是遇到的第二个问题。

为了解决以上两个问题，本文提出了一种基于 Transformer 的服务推荐方法(SRT)，首先，我们使用 Transformer 来对开发者提出的开发需求进行文本特征提取，充分挖掘开发者需求的隐含语义信息，接着我们使用深度神经网络来进一步挖掘应用和服务的潜在关系，进而进行服务推荐。

## 2. 相关工作

服务推荐是指根据用户需求和偏好，为用户推荐最合适的服务。在过去的研究中，已经有许多关于服务推荐的工作[1]-[5]。

### 2.1. 基于内容的服务推荐方法

基于内容的服务推荐是一种常见的推荐方法，它通过分析服务的特征和属性，为用户推荐与其需求相关的服务。在相关工作中，研究者们已经提出了许多基于内容的服务推荐算法。一种常见的方法是使用关键词匹配。这种方法首先对服务的描述文本进行处理，提取关键词或特征词。然后，通过比较用户需求与服务的关键词进行匹配，为用户推荐与其需求匹配程度较高的服务。例如，Zhong 等人[6]基于由一组组件 Web API 组成的移动应用程序的功能描述来挖掘 Web API 的客观可靠的功能。挖掘过程主要是通过分析移动应用程序的功能描述和应用程序 API 的组成结构。Hao 等人[7]通过挖掘隐藏在移动应用程序描述中的价值信息来描述 Web API 功能。然后，提出了一种有针对性的重构服务描述(TRSD)方法来帮助 Web API 推荐和移动 APP 开发。

### 2.2. 基于神经网络的服务推荐方法

基于神经网络的服务推荐方法通常利用深度学习技术来挖掘用户行为和服务特征之间的复杂关系，实现个性化的推荐。例如，Yan 等人[8]利用 LightGCN 的协作注意力卷积网络有效地捕获双边信息进行服务推荐。Wei 等人[9]提出了一种时间感知服务推荐方法。不幸的是，很难在软件工程场景中验证这种方法。因此，他们使用产品推荐数据集进行实验。Mezni 等人[10]提出了一种时间感知服务推荐方法，该方法利用时间知识图对用户与服务的交互进行建模。Liu 等人[11]提出了一种捆绑服务推荐方法，该方法旨在解决推荐服务之间的约束问题。这些基于神经网络的服务推荐方法通过深度学习技术的应用，能够更好地挖掘用户和服务之间的复杂关系，实现个性化、精准的推荐，为用户提供更好的推荐体验。

## 3. 基于 Transformer 的服务推荐方法(SRT)

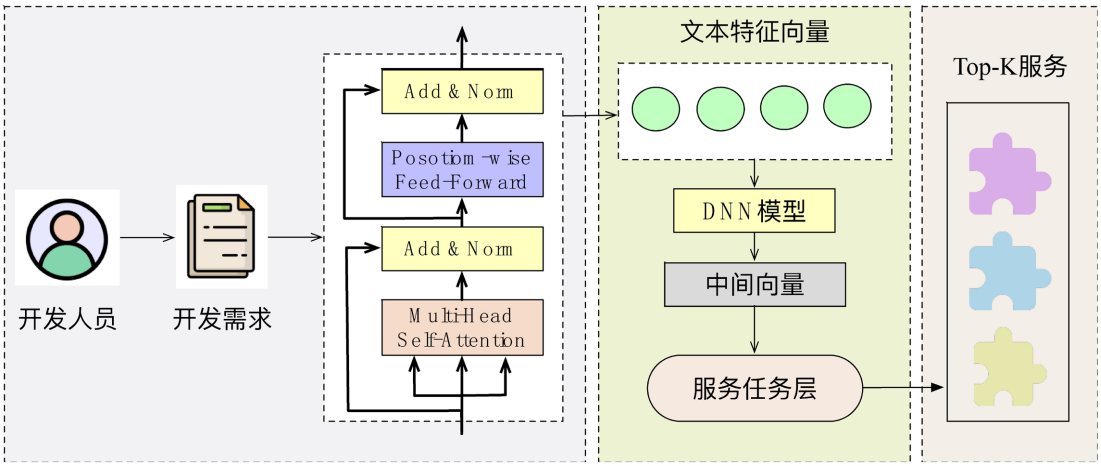


**Figure 2.** The framework of SRT
**图 2.** SRT 框架图

图 2 显示了我们所提出的基于 Transformer 的服务推荐方法的框架。首先，我们使用 Transformer 模型来提取应用的开发需求特征，接着，我们利用深度神经网络 DNN 来进一步挖掘应用和服务的潜在关系，从而进行服务推荐。

### 3.1. 文本特征提取

自从 Transformer [12]被提出以来，这种基于注意力机制的神经网络在许多领域都取得了巨大的成功。Transformer 中的编码器部分在学习单词与单词之间的交互信息方面具有很强的能力。受此启发，本文使用 Transformer 编码器对开发需求进行特征提取。

给定开发需求 $vec_a$，Transformer 编码器对开发需求中的所有单词提取特征表示，输出可表示为 $vec_a = \{vec_1, vec_2, \cdots, vec_L\}$。其中，$L$ 表示开发需求的长度，$vec_i$ 表示在位置 $i$ 的单词的嵌入表示。

编码器由 $N$ 个相同的层组成，每一个层包含了两个子层，分别为自注意机制和完全连接的前馈网络。假设输入为 $g$，子层应用残差连接，然后进行层归一化来计算输出，可以表示为 $LayerNorm\left(g + SubLayer\left(g\right)\right)$，其中 $SubLayer$ 表示自注意机制和完全连接的前馈网络。

由于编码器的每一层都相同，这里选取其中的第 $l$ 层来进行简单介绍。第 $l$ 层的输入为前一层的输出 $vec_a^{l-1}$，第一层的输入为文档的嵌入表示 $vec_a = X$。在自注意机制中，查询 $Q$、关键字 $K$ 和值 $V$ 都是具有不同参数矩阵的 $vec_a^{l-1}$ 的线性投影，自我注意力的输出是根据缩放的点积注意力来计算的，计算过程如下所示：

$$Att\left(Q,K,V\right) = Soft\max\left(\frac{Q^T K}{\sqrt{d}}\right)V \tag{1}$$

此外，Transformer 编码器并行计算 $h$ 头实现自我注意，其中每个头基于公式(1)计算注意力。多头注意力的输出是 $h$ 个头的串联，接着进行线性投影，过程如下：

$$MultiHead\left(Q,K,V\right) = f\left(H_1, H_2, \cdots, H_h\right)W^M \tag{2}$$

$$H_i = ATT\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{3}$$

其中，$f$ 表示拼接操作，$W^M$，$W_i^Q$，$W_i^K$，$W_i^V$ 表示参数矩阵。在自注意机制子层后，完全连接子层获取自注意的输出，表示如下：

$$FFN\left(X^l\right) = \mathrm{Re}\,LU\left(E_{ATT}^l W_1 + b_1\right)W_2 + b_2 \tag{4}$$

其中，$W_1$，$W_2$，$b_1$ 和 $b_2$ 表示可训练参数。

### 3.2. 服务应用关系挖掘

在获取开发需求特征 $vec_a$ 后，接着使用 DNN 层，进一步利用应用和服务的交互记录，提取应用和服务之间的非线性关系。这个过程可以表示为：

$$vf_a = w_i\left(vec_a\right) + b_i \tag{5}$$

在这里，DNN 的优点是可以学习不同抽象级别的交互特性。随着层数的增加，每个神经元的感受野相对于前一层变得更大。通过这种方式，它可以提供全局语义(全局交互)和抽象细节，这在浅层和线性操作中很难做到。

最后，将学习到的向量 $vf_a$ 输入到 sigmoid 函数中，在这里，加入了多任务学习。作为一种归纳转移方法，多任务学习可以充分利用多个相关任务的训练信号中隐含的特定领域信息，从而可以减少过拟合风险。同时，多任务学习有助于模型关注最基本的特征并提高其泛化能力。最后，多任务学习中的辅助

任务所提供的额外信息也有助于学习特征之间的相关性或不相关性，从而帮助模型更好地理解数据。最后输出为 $\hat{y}_c$、$\hat{y}_{ac}$、$\hat{y}_{sc}$。$\hat{y}_c$ 表示表示下一个服务被推荐的概率，$\hat{y}_{ac}$ 表示应用 $a$ 属于某个类别的概率，$\hat{y}_{sc}$ 表示服务 $s$ 属于某个类别的概率。

$$\hat{y}_c = \text{sigmoid}\left(W_c v f_a + b_c\right) \tag{6}$$

$$\hat{y}_{ac} = \text{sigmoid}\left(W_{ac} v f_a + b_{ac}\right) \tag{7}$$

$$\hat{y}_{sc} = \text{sigmoid}\left(W_{sc} v f_a + b_{sc}\right) \tag{8}$$

其中 $W_c$、$W_{ac}$ 和 $W_{sc}$ 是权重矩阵，$b_c$、$b_{ac}$ 和 $b_{sc}$ 是特定任务层中的偏差向量。这里，为了抑制提出的模型的过度拟合，采用了 dropout。在这里，设置 dropout = 0.2。

## 4. 实验

### 4.1. 数据集

2021 年 3 月，从全球最大的在线 Web 服务注册中心 ProgrammableWeb 中获取到 22,016 个服务和 6,438 个应用程序。原始数据集中清除了任何没有功能描述的应用程序和服务、从未使用过的服务以及包含少于三个组件服务的应用程序。最终的实验数据集包括 1,384 个服务和 6,438 个应用程序。数据集的综合统计数据如表 1 所示。使用五重交叉验证技术评估不同的推荐方法。换句话说，数据集被分成了五个部分。每次，一个样本用于测试，另外四个样本用于训练。然后对五次的结果求平均值，并将平均值作为最终的结果。

**Table 1.** Dataset statistics (after preprocessing)
**表 1.** 数据数据集统计(预处理后)

| | 值 |
|---|---|
| #应用程序 | 6438 |
| #服务 | 1384 |
| #应用程序标签 | 417 |
| #服务标签 | 366 |

### 4.2. 评价指标

使用 Precision、Recall、F1 进行性能评估，定义如下公式：

$$\text{Precision}@k = \frac{1}{n}\frac{\left|top_a(k) \cap test_a\right|}{k} \tag{9}$$

$$\text{Recall}@k = \frac{1}{n}\frac{\left|top_a(k) \cap test_a\right|}{\left|test_a\right|} \tag{10}$$

$$\text{F1}@k = \frac{\text{Precision}@k \times \text{Recall}@k}{\text{Precision}@k + \text{Recall}@k} \tag{11}$$

在数据挖掘中，通常将训练集中的数据划分为正样本和负样本，在具体的例子中，正样本是指某应用程序调用的服务，负样本表示是某应用程序没有调用过的服务。$\text{Precision}@k$ 预测为正样本中，被实际为正样本的比例。如公式(9)所示，其中 $top_a(k)$ 推荐给应用程序 $a$ 的前 $k$ 个服务，而 $test_a$ 测试集中应用程序 $a$ 实际调用的服务。$\text{Recall}@k$ 实际为正样本中，预测为正样本的比例。F1 是 Precision，Recall 和

平均数，是一个综合指标，可以更全面地反映服务推荐的性能。通常情况下，F1 越高，代表服务推荐的性能越好。

### 4.3. 对比方法

Word2Vec [13]：Word2Vec 方法首先使用 Word2Vec 工具对开发需求的文本描述特征进行提取。然后基于提取的特征输入到 DNN 模型中计算应用对于各个服务的偏好。

Doc2Vec [14]：Doc2Vec 方法首先使用 Doc2Vec 工具对开发需求的文本描述特征进行提取。然后基于提取的特征输入到 DNN 模型中计算应用对于各个服务的偏好。

### 4.4. 性能分析

图 3 显示了不同方法在精确率、召回率、F1 值方面的性能比较。由于应用程序服务调用记录的密度很低，服务推荐的准确性等指标总体表现不高。从图中可知，Word2Vec、Doc2Vec 等方法无论在精确率还是在召回率等指标上来看，其表现都低于我们所提出的方法。证明了相对于传统的特征提取方法，Transformer 方法在文本特征提取方面显示出了优势。与次优方法相比，SRT 在 Precision@1 上提高了 15.2%，在 Recall@1 上提高了 15.1%，在 F1@1 上提高了 15.0%。

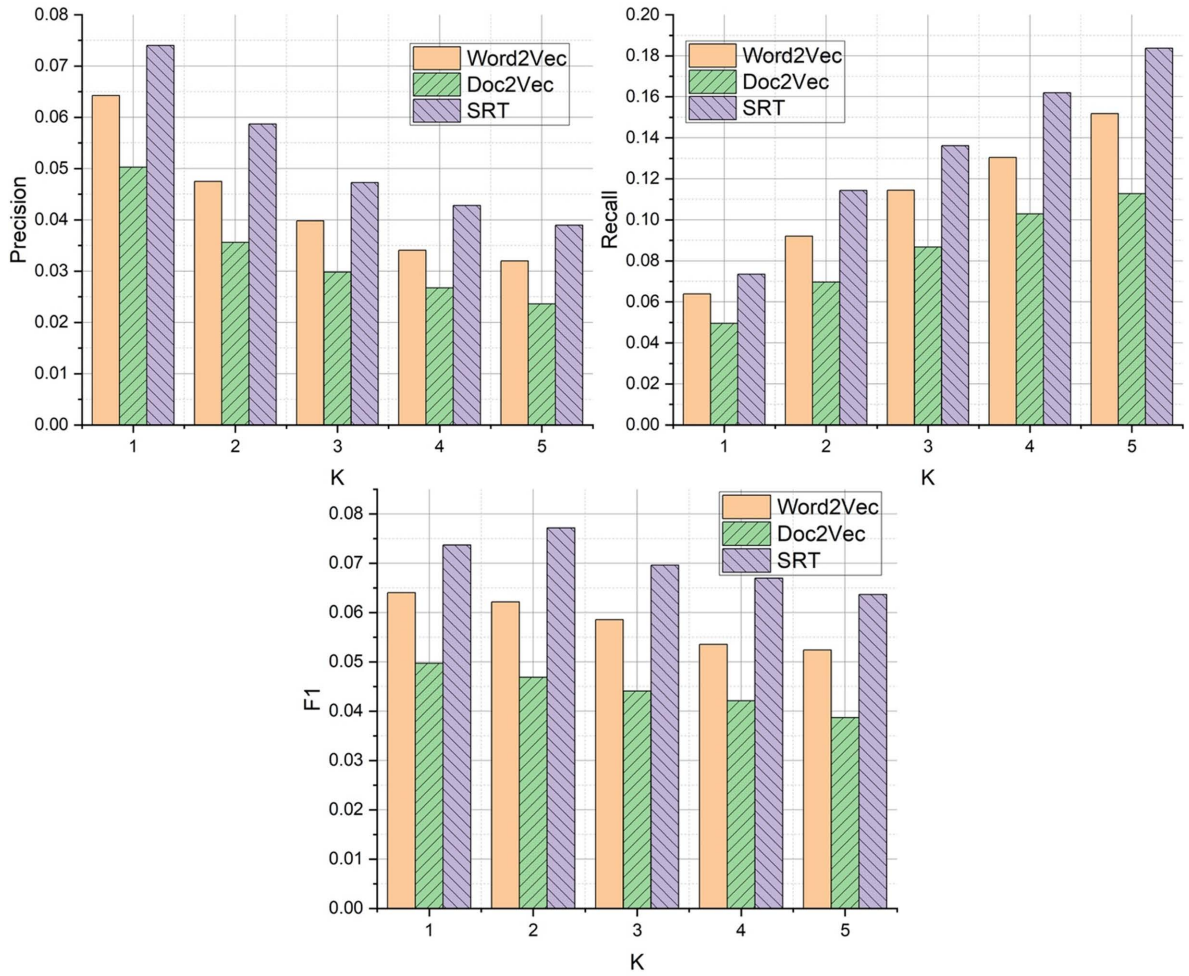


**Figure 3.** Comparison results of different methods
**图 3.** 不同方法对比结果

## 5. 结论和未来工作

Web服务的重用使得软件开发人员可以快速、经济地创建应用程序。本文提出了一种基于Transformer的服务推荐方法SRT。在SRT中，考虑了开发需求的文本信息，通过DNN模型和多任务学习来预测候选服务被选择的概率。最后，通过一组来自ProgrammableWeb的真实数据实验验证了SRT的有效性。服务多样性也会影响开发者的满意度，在未来的研究中，除了探索更多的服务兼容性方法外，计划探索挖掘长尾服务，提高服务推荐的多样性和开发者满意度。

## 基金项目

国家级大学生创新创业训练计划(No.202313291024)；嘉兴南湖学院大学生研究训练计划(No.8517233215)。

## 参考文献

[1] 黄沈权, 朱晓辉, 陈子瑞, 等. 基于双向序列特征和主题语义模型的制造服务推荐方法[J]. 计算机集成制造系统, 2024: 1-28.

[2] 黄德玲, 童夏龙, 杨皓栋. 融合图注意力网络和注意力因子分解机的服务推荐方法[J]. 重庆邮电大学学报(自然科学版), 2024, 36(2): 357-366.

[3] 刘庆雪, 王荔芳, 潘国庆, 等. 面向功能语义增强与标签关联的 Web 服务标签推荐[J]. 计算机应用研究, 2024: 1-8.

[4] 刘佳慧, 袁卫华, 曹家伟, 等. 基于用户特征聚类与服务质量预测的推荐方法[J]. 南京大学学报(自然科学), 2023, 59(1): 120-133.

[5] 杨洁, 朱咸军, 周献中, 等. 基于混杂社会网络的个性化 Web 服务推荐方法[J]. 电子学报, 2020, 48(2): 341-349.

[6] Zhong, Y., Fan, Y., Tan, W. and Zhang, J. (2018) Web Service Recommendation with Reconstructed Profile from Mashup Descriptions. *IEEE Transactions on Automation Science and Engineering*, **15**, 468-478. https://doi.org/10.1109/tase.2016.2624310

[7] Hao, Y., Fan, Y., Tan, W. and Zhang, J. (2017). Service Recommendation Based on Targeted Reconstruction of Service Descriptions. 2017 *IEEE International Conference on Web Services* (*ICWS*), Honolulu, 25-30 June 2017, 285-292. https://doi.org/10.1109/icws.2017.44

[8] Yan, R., Fan, Y., Zhang, J., Zhang, J. and Lin, H. (2021). Service Recommendation for Composition Creation Based on Collaborative Attention Convolutional Network. 2021 *IEEE International Conference on Web Services* (*ICWS*), Chicago, 5-10 September 2021, 397-405. https://doi.org/10.1109/icws53863.2021.00059

[9] Wei, C., Fan, Y. and Zhang, J. (2022) Time-Aware Service Recommendation with Social-Powered Graph Hierarchical Attention Network. *IEEE Transactions on Services Computing*, **16**, 2229-2240. https://doi.org/10.1109/tsc.2022.3197655

[10] Mezni, H. (2022) Temporal Knowledge Graph Embedding for Effective Service Recommendation. *IEEE Transactions on Services Computing*, **15**, 3077-3088. https://doi.org/10.1109/tsc.2021.3075053

[11] Liu, M., Tu, Z., Xu, H., Xu, X. and Wang, Z. (2023) DySR: A Dynamic Graph Neural Network Based Service Bundle Recommendation Model for Mashup Creation. *IEEE Transactions on Services Computing*, **16**, 2592-2605. https://doi.org/10.1109/tsc.2023.3234293

[12] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[13] Mikolov, T., Chen, K., Corrado, G., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space.

[14] Le, Q. and Mikolov, T. (2014) Distributed Representations of Sentences and Documents. *International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 1188-1196.