知识自动化方法初探：从交通信号控制和互联网问诊平台

赵一飞
中国科学院自动化研究所
复杂系统管理与控制国家重点实验室
北京，中国
yfyf.zhao@163.com

施小博
青岛智能产业技术研究院
智慧健康所
青岛，中国
xiaobo.shi@qaii.ac.cn

***摘要*—基于大数据的知识自动化是开发人工世界的核心支撑科学和技术。以麦肯锡《颠覆技术：即将变革生活、商业和全球经济的进展》报告中预测知识工作的自动化技术的重要性为一个标志性事件，国内外很多学者从不同领域对知识自动化进行了相关研究，并提出了相应的方法论。本文尝试从交通信号控制（物理过程）和互联网问诊平台（信息过程）两个领域，分析其不确定性、多样性、复杂性的特点，落实知识自动化在这两个领域的具体方法。**

关键词—知识自动化；交通信号控制；互联网问诊；方法

I. 引言

2013 年麦肯锡全球研究所在其报告《颠覆技术：即将变革生活、商业和全球经济的进展》[1] 中预测了 12 项可能在 2025 年之前决定未来经济的颠覆性技术，其中知识工作的自动化智能软件系统位居第二。以此为标志性事件，国内外专家学者对知识自动化进行了相应的解读，提出相关方法论，并深化到相关领域[2][3][4][5][6]。

文献[7]认为，如何科学、准确地界定知识自动化的内容，目前实际不成熟也没有必要。但在很大程度上，知识自动化可以狭义地理解为基于知识的服务（Knowledge-based services, KBS），与基于位置的服务（Location-based services, LBS）等类似。其关键点在于，如何把知识与任务或需求无缝、准确、及时、在线地结合起来。广义的理解知识自动化，可以粗略地认为是一种以自动化的方式变个性地改变知识产生、获取、分析、影响、实施的有效途径。

方法论层面上，文献[7]指出，由于知识自动化的研究的主要动机是面向复杂系统，解决复杂问题，其最迫切的任务是如何将复杂系统的“不确定性、多样性、复杂性”（UDC: Uncertainty, diversity, complexity）等特征转化为智能系统的“灵捷、聚焦、收敛”(AFC: Agile, focus, convergence)特性，所以知识自动化需要嵌入到基于 ACP 的平行控制与管理的框架和流程之中，使复杂变为简单，使 UDC 化为 AFC。具体而言，ACP 分三步：首先将复杂问题建模成人工社会或人工系统；然后利用计算实验对复杂现象进行分析和评估；最后将实际系统和人工系统实现虚实互动，以平行执行的方式对复杂系统的运行进行有效地控制和管理[8][9]。

基于以上认知和方法论，知识自动化已在工业和智能制造 [4]、智能指挥与控制 [3]、情报[2]、智能交通[5]、决策管理[10]领域得到深化。

由此可见，知识自动化不是单一具体的方法。在前人的基础上，本文进一步具体落实其在交通信号控制和互联网问诊平台上的具体方法。数据驱动之前的交通信号控制，属于典型的传统控制领域，是实实在在的物理过程，主要是捕捉各种相关信号，对过程进行精确地建模，再实施控制，落实设定的目标；但其在建模时会遇到很多不确定的、难以描述和准确量化的因素，导致理论上的理想模型难以在实际中达到理想中的效果。大数据之前的互联网问诊平台，其特色功能有相似病例自动推荐，其达到信息过程自动化程度；但由于描述的多样性和碎片化，导致无法从语义角度去真正匹配相似病例，真正的自动化程度不够高。本文根据之前的工作，已对交通信号控制实现知识自动化，对互联网问诊平台提高信息自动化程度，针对过程中遇到的海量、难以准确描述、碎片化的知识，结合领域特点分析，探寻出哪些具体方法可以实现知识自动化，这是本文的主要贡献。

II. 交通信号控制

交通信号控制（Traffic Signal Control, TSC）被认为是一种有效的缓解交通拥堵问题的方法，并且得到了越来越多的研究者的重视 [11][12][13][14][15]。一般来讲，各种各样的城市交通信号控制策略可大致分为如下四类：定时控制、感应控制、自适应控制和智能控制。定时控制策略的周期和绿信比都是预先设定的，具有实施方便的优点，比较适用于交通流平稳和规则的情况 [16][17]。感应控制可以取得比定时控制更好的效果，因为其信号配时可以随着交通流的变化而改变。但是为了获取实时的交通流信息，我们必须花费很多金钱和时间去购买和安装检测器。

另外，当交通流剧烈变化和车道饱和度达到 80%左右时，感应控制并不能表现出很好的性能。自适应控制将交通系统看做一个不确定性系统，信号配时根据交通状态的反馈信息进行动态的优化[18][19]。除了以上三种策略，智能控制在交通信号控制中越来越流行，如基于模糊逻辑的方法[20][21][22][23]和多代理技术[24][25][26][27]。模糊逻辑提供了一种类人思考的能力，而多代理技术则将所有事物看做一个代理。

以上相关方法已经在一定程度上体现了自动化、智能化，尤其是智能控制策略。然而，尽管这些方法、模型、策略在某种程度上能反映交通状况，但我们必须承认相对于真实复杂的交通环境而言还显得很粗糙。动态的交通环境具有很强的随机性、非线性和复杂性，这导致交通信号控制存在一个固有的问题：很难建立能反映真实交通环境的精确的数学模型。因为交通环境中有很多难以捕捉、难以描述、难以量化的不确定性因素。同时，按这种传统思路首先就需要结合各种因素对复杂的交通环境进行建模，这会存在另外两个问题：1）尽管有些模型通过人工极力考虑各种因素，建立在某种交通环境下尽可能准确的数学模型，但这些模型往往复杂和费时；2）复杂的交通环境多变，理论上尚不存在一种通用的方法适用于各种交通环境。所以，我们必须结合知识自动化重新审视交通信号控制，从一个全新的角度寻求新的突破。

文章[28]指出：知识自动化绝对不是知识本身的自动产生，但可以诱发知识的传播、获取、分析、影响、产生等方面的重要变革。交通信号控制恰好是对交通环境中的各种因素（知识）进行获取、分析、影响的过程，其关键阻碍在于无法自动获取全面的因素、难以精确分析可人工获取的因素对结果的影响。但全面获取因素目的是有利于更精确的分析交通环境，精确分析的目的又是考虑其影响程度以便于实施控制，所以我们的根本目的在于能够自动感知其影响。同时，文章[7]又指出，电子商务中获得成功应用的各种推荐系统可以在知识自动化中发挥重要而有效的作用。结合 ACP 方法，我们设计出图 1 所示的交通信号控制自动化的推荐系统。

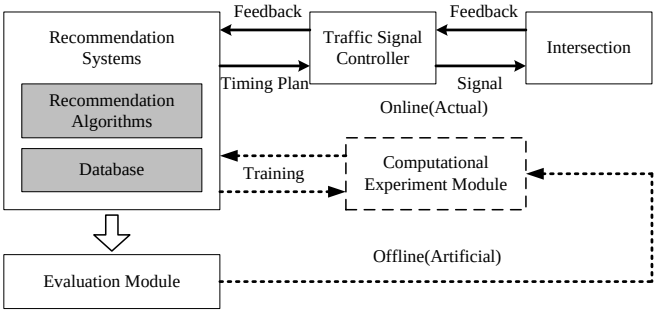

图 1

该系统从一个全新的角度，将传统的根据各因素配置信号配时，转换成从配时方案库搜索进而进行自动推荐。该系统的前提是，每次每刻都有大量的交通数据产生，涉及需求、配时方案、以及可反映前两者匹配度的不断的变化的路况。在有交通大数据的前提下，我们则可以挖掘出各种交通因素的内在联系，从而将交通需求看做推荐系统里的用户，配时方案看做被推荐的物品，相应的路况可反映配时方案能满足当前需求的程度，可看做推荐系统中的打分。该系统的设计主要出于如下考虑：1）类似于无模型自适应思想，该系统在建模时无需具体考虑交通环境中各种因素，但其影响效果却不会被忽略；2）该系统可离线训练学习，结合当前路况在线推荐，同时保证准确率和实时性；3）推荐系统从全局角度寻求与当前路况最匹配的配时方案；4）该系统具有自适应和自完善能力：系统会根据每次推荐的反馈结果不断的更新和完善数据库。具体而言，该系统由在线和离线两部分组成，离线部分是虚拟的，用于预测，在线部分是真实的，用于推荐。通过计算实验方法，虚拟部分利用各种推荐算法预测数据库中交通需求和配时方案的匹配程度，然后结合实际的交通情况推荐合适的配时方案。在实际采用被推荐的配时方案后，系统会记录最新的反馈信息，以进一步完善或更新数据库。所以，在假定已经建立了一个庞大的数据库的前提下，该系统的核心就在于用于预测和推荐的各种推荐算法。

用于推荐系统最流行的是协同过滤的一系列算法。我们在文章[29]中，提出推荐代理的概念，以交通需求 s、配时方案 t、相应的打分 r 组成三元数组（s，t，r）作为推荐代理，通过协同过滤的思想进行推荐。具体而言，我们可以得到很多交通需求 s，很多配时方案 t，以及部分相应打分 r，我们的任务就是要利用已有的数据，预测 s 对未曾使用过的 t 的打分情况，进而推荐出最能满足当前交通需求 s 的配时方案 t。进一步可发现，该三元数组可转换成一个矩阵，其列对应交通需求、行对应配时方案，而矩阵元素则对应打分；由于部分打分是需要预测的，所以可以转换成矩阵填充问题。这在数学上有很多方法，结合实际情况，我们在文中采用了 slope-one 和 weighted slope-one 算法。这样便实现了配时方案的自动推荐。同样，我们在文章[30]中，采用隐语义模型来预测 s 对未知 t 的打分。我们认为配时方案和交通需求的匹配度是和交通环境中的各种因素相联系的，这些因素即那些难以捕捉、难以描述、难以量化的因素。通过隐语义模型，我们利用这些因素建模交通需求和配时方案，但无需具体知道和量化这些因素，通过大量数据，采用机器学习算法，如 SVD 将其学习出并预测。这样便实现了对处在复杂交通环境中的交通需求和配时方案的自动建模，进而可进行预测和自动推荐。

以上两种方法都存在一个问题，即当一个新的交通需求出现时，无法直接从数据库中寻找到相应的数据对其进行挖掘。在复杂多变的交通环境中随处可见，这便是冷启动问题。对于这个问题，我们在文章[31]中基于交通要素的相关特征，对交通需求采用用户画像技术，从内容层面进行预测和推荐，同样也达到了相应的效果。这种方法与上述方法结合，便可全面的实现交通信号控制中配时方案的自动推荐。

我们从文中选取两个实验效果图进行分析。图 2 是我们对 40 中交通需求进行配时方案推荐，根据预测结果给出的推荐列表和实际最优排序列表的一个契合度指标 nDCG [32]，其值最大为 1，约接近 1 表示契合度越大，为

1 时表示推荐列表和实际列表完全一致。从图中可以看出其整体效果很好。图 3 是将我们的系统效果和单一方法做一个对比，可明显发现其优越性；并且传统单一方法，无法适用于所有交通需求，所以在交通需求 1,2,3,4 时，传统单一方法没有结果，但我们的自动推荐系统则有较好的表现。同时，如果单一方法在某些情况下优于系统性能，系统则会吸收该方法，以进一步优化和完善自身 [31] 。综上所述，可证明通过我们的方法在交通信号控制领域实现知识自动化的可行性和优越性。

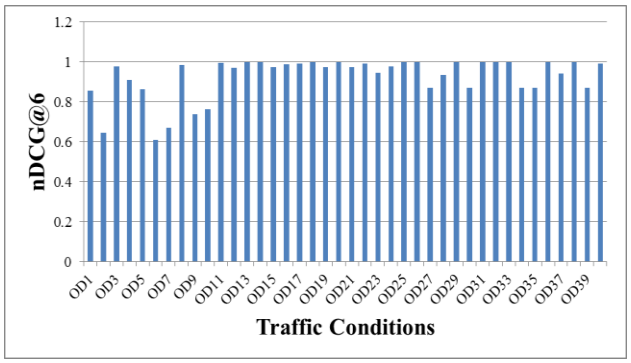

图 2

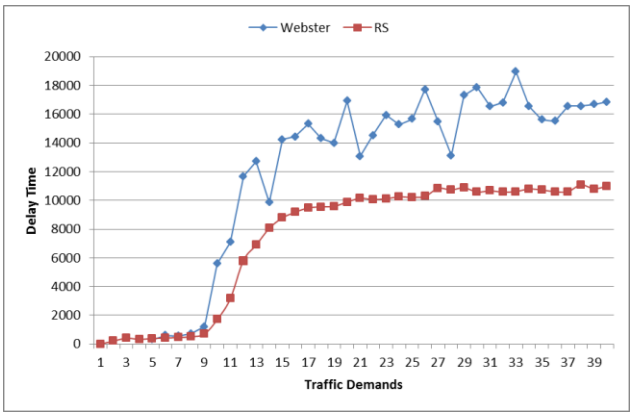

图 3

III. 互联网问诊平台

互联网问诊近年来作为传统医疗服务体系的补充和优化越来越流行，如 39 问医生[33]、好大夫在线[34]、春雨医生[35]等。其中一个最重要的功能就是在线问诊，即平台的一方是患者，平台的另外一方是医生；患者描述他们的身体状况，医生在线给出专业性建议。为了避免患者过长时间的等待，绝大部分问诊平台都提供相似病例推荐功能。由于平台上已积累了大量已高质量解决的病例，相似病例推荐的推荐不只是可能，而且对于资源优化是一个必要性功能。

相似病例推荐的核心任务，给定身体状况的描述，从语义层面在大量的历史病例库中找到相似的可供参考的病例。那些有最佳答案的相似病例可以作为查询问题的较好的参考。所以，任务就被转换成计算当前病症描述和历史病例之间的相似性问题。很多学者对该问题进行了相关研究[36][37][38][39][40][41][42][43]。然而，对于词汇鸿沟问题[44][45]，传统方法因为过分依赖于词的全匹配而无法取得很好的效果。例如，“如何减肥”的最佳答案可能存在于“怎么瘦身”的病例中，尽管两句话描述的是同一个意思，但却没有任何共同词汇。传统的方法，如向量空间模型[46]，Okapi BM25[47]，查询似然语言模型[48]等，都要求两个句子之间有一定的共现词语。他们缺乏语义层面的挖掘能力，从而无法应对随处可见的词汇鸿沟现象。

在过去几年中，也有很多学者尝试通过一些模型对不同描述的词汇进行“自动认知”，其中效果较为明显是基于统计翻译的方法[36][37][38][39][40][41][42][43]。这些方法假定问答对为“平行语料库”，用来估计词与词、短语与短语之间的翻译概率，而翻译概率又可看做语义相关性来提升传统检索模型。这些方法在一定程度上缓解了“词汇鸿沟”问题，并且实验结果显示优于传统的方法[36][49]。然而，尽管基于统计翻译的模型可以从语义的角度找到两个相似的词语，但实际上问答对并不能真正作为“平行语料库”。他们不仅在对称性上存在较大差异，而且完全是从两个不同的角度描述问题。

受最近在各种自然语言处理任务中表现较佳的词嵌入的启发[50][51]，我们将它嵌入检索模型以期更好的推荐。词嵌入通过将词映射到一个连续空间抓取语义。基于词嵌入，我们主要做了三方面的工作，1）通过词嵌入解决词汇鸿沟问题，并将其嵌入到查询似然语言模型中，获得了更好的相似病例推荐；2）直接将病例映射到连续空间，然后和词嵌入一起整合进查询似然语言模型，或直接计算病例相似度；3）结合医学数据特点，尝试一些策略自动加强相关医学词汇的权重，获得更好的推荐。具体而言，我们在文章[52]中，首先将句子相似性计算转换成生成概率，生成概率的计算再分解成查询似然语言模型，然后再用词嵌入代替查询语言模型中的似然估计，最终可从语义层面获得两个病例的相似性。其过程如下式（1）~式（5）所示，

$$sim(Q,D) \approx P(Q|D) = \prod_{\omega \in Q} P(w|D) \quad (1)$$

$$P(w|D) = (1-\lambda) P_{ml}(w|D) + \lambda P_{ml}(w|Conll) \quad (2)$$

$$P_{ml}(w|D) = \frac{\#(w,D)}{|D|} \quad (3)$$

$$P_{ml}(w|Conll) = \frac{\#(w,Conll)}{|Conll|} \quad (4)$$

$$P(w|D) = (1-\lambda) \sum_{t \in D} sim(w,t) P_{ml}(t|D) + \lambda P_{ml}(w|Conll) \quad (5)$$

其中，式（1）为相似性转概率，式（2）为概率转查询似然计算，式（3）（4）为原始似然计数计算法，式（5）为词嵌入结合后的似然计算。

在以上基础上，我们做了相关变种，将词汇鸿沟从词解决的角度扩展至句子层面和策略层面，共做了 5 中模型，并且与之前较好的基于翻译的两种方法做了对比，其结果如图 4 所示。

Models	Precision@20			MAP
	1	2	3	
WTM	10.55%	14.57%	74.87%	58.61%
TRLM	11.17%	9.14%	79.70%	66.92%
WEM	9.55%	7.04%	83.42%	73.86%
WELM	9.05%	7.54%	83.42%	75.74%
PVM	13.00%	6.00%	81.00%	72.30%
PVLM	6.5%	7.5%	86.00%	76.09%
WAMM	9.14%	8.63%	82.32%	73.97%

图 4

其中，WTM 和 TRLM 是两种经典的基于统计翻译的模型，其余为我们提升知识认知自动化程度的模型，准确率中的 3 表示最相似，1 表示最不相似。从结果可以看出，我们的方法由于基于统计翻译的方法，能更好的从语义层面进行自动认知。

IV. 总结和展望

本文从物理过程和信息过程各选取一个领域，物理过程的交通信号控制和信息过程的互联网问诊，初探了可实现知识自动化的方法。总结而言，主要是两种方法或技术：协同过滤和词嵌入。

仔细分析，对于物理过程中大量的难捕捉、难描述、难量化的因素，在有大数据的情况下，可通过协同过滤方法，利用相似群体的信息，以结果为导向，建模时忽略其因素本身却保证不丢弃其影响。从这一层面来讲，可以对很多物理过程的情况利用协同过滤实现知识自动化，而无需过多的人工介入，去收集信息、处理信息、分析信息、建模信息、实施控制。对于信息过程，由于描述的多样性，一般算法很难自动智能认知；在大数据量的情况下，很多机器学习算法都表现出了一定的自动化特性，但一般研究中监督学习居多，而监督学习则需要标记，现实生活中的大部分数据又是没有标记的。词嵌入无需标记，从表层看是一种无监督学习；深入洞察其原理可发现，其实词嵌入是一种自监督的学习过程。因为大量的文本，大量的上下文信息，和独特的训练技巧，使得它可利用自身的信息作为标记进行内在自监督外在无监督的学习，从而实现了互联网问诊平台的知识自动化，其方法、思想可尝试用于实现其他更多信息过程的知识自动化。

知识自动化方法不是单一的方法，我们接下来的工作是从当前工作收到的启发，继续寻找新的领域，尝试进一步落实知识自动化。

参考文献

[1] McKinsey Global Institute. Disruptive technologies: advances that will transform life, business, and the global economy [Online], available: http://www.mckinsey.com, June 6, 2013

[2] 王飞跃. 情报 5.0: 平行时代的平行情报体系[J]. 情报学报, 2015, 34(6): 563-574.

[3] 王飞跃. 指控 5.0: 平行时代的智能指挥与控制体系[J]. 指挥与控制学报, 1(1): 107-120.

[4] 王飞跃. 平行工业 5.0: 平时时代的智能制造体系. 见: 2015 年国家智能制造新年论坛. 北京, 2015.

[5] Wang F Y. Scanning the issue and beyond: Toward ITS knowledge automation[J]. Intelligent Transportation Systems, IEEE Transactions on, 2014, 15(1): 1-5.

[6] 王飞跃. 从牛顿系统到默顿系统——系统工程的过去, 现在和未来: 系统工程与管理变革: 从牛顿到默顿的升华[J]. 管理学家: 实践版, 2013 (10): 10-19.

[7] 王飞跃. 软件定义的系统与知识自动化: 从牛顿到默顿的平行升华[J]. 自动化学报, 2015, 41(1): 1-8.

[8] Fei-Yue W. Parallel system methods for management and control of complex systems [J][J]. Control and Decision, 2004, 5: 001.

[9] Wang F Y. Parallel control: A method for data-driven and computational control[J]. Acta Autom Sin, 2013, 39: 293-302.

[10] Fish A N. Knowledge Automation: how to implement decision management in business processes[M]. John Wiley & Sons, 2012.

[11] Z. Liu, "A survey of intelligence methods in urban traffic signal control," *IJCSNS International Journal of Computer Science and Network Security*, vol. 7, no. 7, pp. 105–112, July 2007.

[12] D. Zhao, Y. Dai, and Z. Zhang, "Computational intelligence in urban traffic signal control: A survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. PP, no. 99, pp. 1 –10, 2011.

[13] C. Chen, F.-H. Zhu, and Y.-F. Ai, "A survey of urban traffic signal control for agent recommendation system," in *2012 15th International IEEE Conference on Intelligent Transportation Systems*, Anchorage, AK, 2012, pp. 327-333.

[14] M. Smith, "Traffic signal control and route choice: A new assignment and control model which designs signal timings," *Transportation Research Part C: Emerging Technologies*, vol. 58, Part C, pp. 451-473, Sept. 2015.

[15] I. Yang, R. Jayakrishnan, "Real-time network-wide traffic signal optimization considering long-term green ratios based on expected route flows," *Transportation Research Part C: Emerging Technologies*, vol. 60, pp. 241-257, Nov. 2015.

[16] F. V. Webster. (1958). Traffic signal settings. *Great Britain Department of Scientific A.* [Online]. 39. Available: www.bl.uk/services/document/lps.html

[17] A. J. Miller, "Settings for fixed-cycle traffic signals," *Oper. Res. Q.*, vol. 14, no. 4, pp. 373–386, 1963.

[18] Liu D R, Zhang Y, Zhang H G. "A self-learning call admission control scheme for CDMA cellular networks." *IEEE Transactions on Neural Networks*, 2005, 16(5): 1219−1228

[19] Sims A G, Dobinson K W. "The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits". *Vehicular Technology, IEEE Transactions on*, 1980, 29(2): 130-137.

[20] C. Pappis and E. Mamdani, "A fuzzy logic controller for a traffic junction," *IEEE Trans. Syst., Man Cybern.*, vol. 7, no. 10, pp. 707–717, Oct. 1977.

[21] M. Trabia, M. Kaseko, and A. Murali, "A two-stage fuzzy logic controller for traffic signals," *Transp. Res. C, Emerg. Technol.*, vol. 7, no. 6, pp. 353‑367, 1999.

[22] J. H. Lee and K. H. Lee, "Distributed and cooperative fuzzy controllers for traffic intersection group," *IEEE Trans. Syst., Man Cybern. C, Appl. Rev.*, vol. 29, no. 2, pp. 263‑271, May 1999.

[23] I. Kosonen, "Multi-agent fuzzy signal control based on real-time simulation," *Transp. Res. C, Emerg. Technol.*, vol. 11, no. 5, pp. 389‑403, 2003.

[24] M. Wiering, J. Vreeken, J. Veenen, and A. Koopman, "Simulation and optimization of traffic in a city," in *Proc. IEEE Intell. Veh. Symp.*, 2004, pp. 453‑458

[25] E. D. Ferreira and P. K. Khosla, "Multi-agent cooperation using distributed value functions," in *Proc. IEEE Intell. Veh. Symp.*, 2000, pp. 404-409

[26] M. C. Choy and D. Srinivasan, R. L. Cheu, "Cooperative, hybrid agent architecture for real-time traffic control system," *IEEE Trans. Syst. Man Cybern. A, Syst. Humans*, vol. 33, no. 5, pp. 597-607, Sep. 2003

[27] D. A. Roozemond, "Using intelligent agents for pro-active real-time urban intersection control," *Eur. J. Oper. Res.*, vol. 131, pp. 293-301, 2001

[28] 王飞跃. 天命唯新: 迈向知识自动化——《自动化学报》创刊 50 周年专刊序[J]. 自动化学报, 2013, 39(11): 1741-1743.

[29] Zhao Y F, Kong Q J, Gao H, et al. Parallel management for traffic signal control[C]//Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on. IEEE, 2014: 2888-2893.

[30] Zhao Y F, Gao H, Lv Y S, et al. Latent factor model for traffic signal control[C]//Service Operations and Logistics, and Informatics (SOLI), 2014 IEEE International Conference on. IEEE, 2014: 227-232.

[31] Zhao Y F, Wang F.-Y, Gao H, Zhu F H, Ye P J, et al. Content-based recommendation for traffic signal control[C]//Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on. IEEE.

[32] https://en.wikipedia.org/wiki/Discounted_cumulative_ga

[33] http://ask.39.net/

[34] http://www.haodf.com/

[35] http://www.chunyuyisheng.com/

[36] J. Jeon, W. B. Croft, and J. H. Lee, 2005a. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 84-90.

[37] X. Xue, J. Jeon, and W. B. Croft, 2008. Retrieval models for question and answer archives. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 475-482.

[38] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, 2008. Searching Questions by Identifying Question Topic and Question Focus. In *ACL (2008)*, 156-164.

[39] J.-T. Lee, S.-B. Kim, Y.-I. Song and H.-C. Rim, 2008. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 410-418.

[40] Bernhard, Delphine, and I. Gurevych, 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In P*roceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*: Volume 2 - Volume 2Association for Computational Linguistics, 728-736.

[41] X. Cao, G. Cong, B. Cui, and C. S. Jensen, 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of the 19th international conference on World wide web*, 201-210.

[42] Zhou, Guangyou, et al. 2011. Phrase-based translation model for question retrieval in community question answer archives. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1Association for Computational Linguistics, 653-662.

[43] Zhang et al, 2014. MapReduce-based Approach on Short Text Conversation Clustering. In *Journal of Computational Information Systems* 10: 8 (2014), 3511–3521

[44] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 192-199.

[45] A. Berger and J. Lafferty, 1999. Information retrieval as statistical translation. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 222-229.

[46] G. Salton, A. Wong, and C.-S. Yang, 1975. A vector space model for automatic indexing. In *Communications of the ACM*, vol. 18, 613-620.

[47] A. Singhal, G. Salton, M. Mitra, and C. Buckley, 1996. Document length normalization. In *Information Processing & Management*, vol. 32, 619-633.

[48] C. Zhai and J. Lafferty, 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 334-342.

[49] J. Jeon, W. B. Croft, and J. H. Lee, 2005b. Finding semantically similar questions based on their answers. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 617-618.

[50] Mikolov T, Chen K, Corrado G, et al, 2013. Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781.

[51] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, September 26-30, 1045-1048.

[52] Zhao Y F, Wang J, Wang F.-Y, Shi X B, Word Embedding based retrieval models for similar cases recommendation. 中国自动化大会, 2015