Latent Relevance

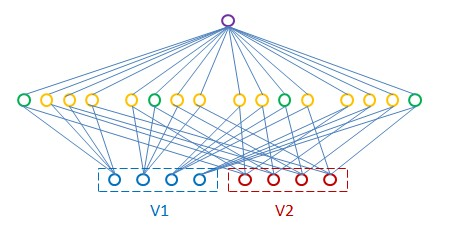
**背景**

此前的调研中发现，公司现有的相关性算法（topic model，word2vec，clicksim等）难以较好地建模不同事物之间的联系，例如“宝马-贷款”，“礼物-香奈儿”，“手机-京东”等等。而这种联系，对于广告来讲，很多时候具有重要意义。这里希望调研一种改进的相关性算法，能够建模不同类事物之间的关系，得到与人脑相关性判断更为接近的结果。 理想情况下，通过共现或共点击，很多相关性算法能够自动学到不同类事物之间的联系。然而，实际中，由于数据、初始化、优化算法、归一化等环节的影响，往往无法达到这种预期。需要对不同事物/不同topic之间的联系，进行显示建模。

另一方面，通用的vector-based model（topic model，word2vec等）大都基于无监督学习，或者利用通用标注数据进行监督/半监督学习得到。而特定应用场景下的标注数据，往往难以有效利用。通常情况下，很多场景能够获得的标注数据量很有限（大多来自于人工标注，例如广告bad case识别，query-langding page相关性判断等）。因此，如何设计模型使之具备更好的泛化性，在小样本场景下能够建模不同事物之间的联系，便成为非常重要且需求广泛的问题。

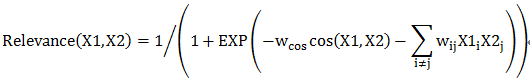
**Latent Relevance**

针对vector-based model，假设vector的每个维度代表一个topic/一类事物，如下模型显式建模了不同topic之间的联系。由于模型引入了隐变量，我们称之为latent relevance model。



图中V1、V2表示两个vector。相比直接计算cosine的区别是，不同维度/topic之间也进行运算，而不仅仅计算对应维度的值，以此来表示不同topic之间的联系。

一种简单的形式（线性）如下：



http://wiki.baidu.com/download/attachments/86029391/a7c5cd22b881a51f4c859d812a184288.png?version=1&modificationDate=1429930802440&api=v2反映了不同topic之间的联系。通过选取合适的训练数据、loss function和正则项，优化求解模型参数，即可得到latent relevance model。

公式中，强制X1、X2对应位按照cosine进行计算，是为了保证模型效果，避免在小数据上过拟合。在乱买词识别问题中（约5000训练数据，其中约1000乱买词，基于PLSA向量构建latent relevance），增加此限制使得测试集AUC提升约1.5%。

线性形式的latent relevance，可以一定程度上理解为广义的马氏距离（Mahalanobis distance）。该方法通过在特定数据下学习马氏距离计算方法，来提升该场景下的相关性度量效果。可以认为，latent relevance的本质不是学习向量表示，而是学习指定场景的距离度量，以此提升模型效果。

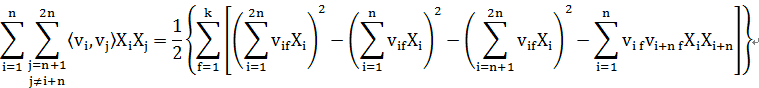
注意：该方法不适用于基于cosine训练出来的词向量表示。因为对该词向量来讲，cosine基本就是最佳距离度量。

另外，值得注意的是，latent relevance是非对称的，即 http://wiki.baidu.com/download/attachments/86029391/36284f01c8fe8c09f1fe3cc98e5c5483.png?version=1&modificationDate=1429930802440&api=v2（当然也可以通过限定Wij=Wji得到对称的模型）。这是by design的，感兴趣的可以参考 [条件相关性](http://wiki.baidu.com/pages/viewpage.action?pageId=86029397)

**Factorization**

建立两两topic之间的联系，使得参数数量大幅膨胀，很容易过拟合（实际应用中，标注数据量可能很小）。并且导致训练、预估计算量显著增加，并不实用。

对于sparse model（PLSA等），基本不存在上述问题，可以直接应用。而对于dense model（word2vec等），则必须对模型进行改进、优化。一个直接的方法是进行factorization，即令 http://wiki.baidu.com/download/attachments/86029391/b2876337c407a6de4c2096f195151b28.png?version=1&modificationDate=1429930802440&api=v2。将模型复杂度从平方降为线性复杂度。这里省略推导，直接给出计算公式（为了表示方便，将X1、X2串联为一个向量X，X1、X2维度均为n）：



将其代入latent relevance计算公式，针对具体问题优化求解即可。

**结论**

一些初步实验结论如下：

1. 针对乱买词识别（约5000训练数据，其中约1000乱买词，基于PLSA向量构建latent relevance），相比PLSA cosine，在测试集AUC约提升4%
2. 针对乱买词识别（约5000训练数据，其中约1000乱买词，基于word2vec向量构建latent relevance，factorization维度设为3），相比word2vec cosine，在测试集AUC约提升11%
3. 针对相关性模型（RIG，约3万训练数据，基于PLSA向量构建latent relevance），相比PLSA cosine，在测试集AUC约提升9%；最终模型准确率持平（>80%）的条件下，召回率提升约9%
4. 针对自主构建数据（约100万训练数据，基于PLSA向量构建latent relevance，详见下方），相比PLSA cosine，在测试集AUC约提升26%

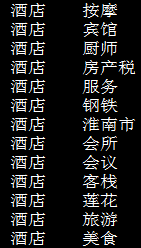
尽管数据不同，不完全可比。但可以认为，一定范围内，训练数据越大，latent relevance的作用越明显。

附件为latent relevance工具，第一版容错可能不够完善，使用中也许会遇到一些问题，工具性能也有待提升（目前适用于10万以下量级的数据）。欢迎感兴趣的同学试用并反馈建议：）

**示例**

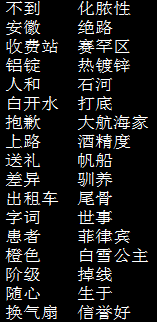
*正样本选取*

关键词的购买蕴含了广告主/客服/SEM的智慧，能够把相关但不同类的事物联系起来，例如平安“贷款”购买“汽车”。实际使用时进行了一些处理，包括仅使用权重较大的term，过滤低clickq数据（防止乱买词），过滤地域（感谢李一的地域词表）、高idf term（客户名称）等等，减少term赋权错误的影响。一些正样本如下：

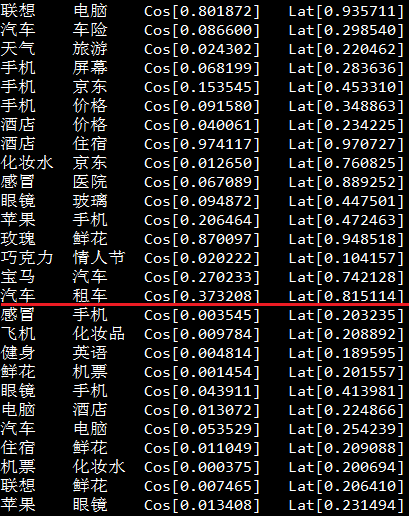


*负样本选取*

负样本通过知道共现term获取，知道问题中共现次数很少的term pair，基本上没有联系，可靠性很高。一些负样本如下：



训练求解模型，得到 latent relevance model，一些与cosine对比的例子如下（红线以上是相关term pairs，红线以下是不相关term pairs）：



可以看出，latent相关性能够提升部分cosine值很低的相关term pair score，同时避免赋予大量不相关term pair较高的score。

基于前面的正负样本选取方法，选取2/3数据用于训练（946786个样本），1/3数据用于测试（473394个样本）。测试集上PLSA cosine [AUC](http://wiki.baidu.com/pages/viewpage.action?pageId=86030888)为0.65，latent relevance[AUC](http://wiki.baidu.com/pages/viewpage.action?pageId=86030888)为0.91，word2vec [AUC](http://wiki.baidu.com/pages/viewpage.action?pageId=86030888)约为0.79。