[Vector-Based非对称相关性](http://wiki.baidu.com/pages/viewpage.action?pageId=86039144)

**背景**

很多常用的字符串相关性算法，都将字符串表示为固定长度的vector（topic model，word2vec及其变种等），通过cosine等方式计算两个字符串的相关性。这类相关性算法都是对称的，即relevance(t1, t2) = relevance(t2, t1)。然而，实际广告应用中，很多时候都是非对称的。例如，用户搜索“感冒”时，关注“医院”的程度，与搜索“医院”时，关注“感冒”的程度明显不同。这些场景下，relevance（d|q）的形式更加适合。

BM25等非对称算法虽然描述了这种形式，但是相比vector-based相关性算法有很多缺点（泛化性差，长串相关性计算不准等），只能互补，不能代替。 这里希望调研一种新的相关性度量方式，针对vector-based的字符串表示，能够衡量relevance（d|q），即给定q，关注d的程度。

当前只考虑相关性计算方式上的改进，如果后续表明可泛化的非对称相关性真的具有重要意义，再进行深入研究。

**Vector-Based非对称相关性计算**

给定两个vector（d, q），计算非对称相关性可以设计很多方法（例如归一化成概率分布，使用KL distance等）。但是很多情况下，可能严重破坏原有相关性的序关系（例如将word2vec归一化成概率分布），得不偿失。因此， 这里希望从理论上推导非对称相关性公式，尽量得到合理的计算方法。下文中relevance简记为Rel，计算目标是Rel(d|q)，后文也称作条件相关性。

首先按照需求对其进行展开（参考概率展开方式，虽然不够严谨，但目前想不到更好的办法），标记基向量

http://wiki.baidu.com/download/attachments/86029397/d4d17f3765cbe868a1624130341a74cd.png?version=1&modificationDate=1429930793739&api=v2

其中 http://wiki.baidu.com/download/attachments/86029397/58a26ae8459c7c132192081b29684666.png?version=1&modificationDate=1429930793739&api=v2与 http://wiki.baidu.com/download/attachments/86029397/8057ed7fa53e640ddd534ff38fe218d1.png?version=1&modificationDate=1429930793739&api=v2符号一致。每个基向量代表一种需求，彼此独立，则

http://wiki.baidu.com/download/attachments/86029397/32f0744b51a33d576d790636dfeaab7c.png?version=1&modificationDate=1429930793740&api=v2

其中

http://wiki.baidu.com/download/attachments/86029397/aca5b0ab218e6af4a66b754e94b6181a.png?version=1&modificationDate=1429930793740&api=v2

其逻辑含义比较清晰，即给定q，关注d的程度（相关性），等于各种需求下关注d的程度乘以q的相应需求强度的总和。 根据 http://wiki.baidu.com/download/attachments/86029397/9027cc00c4601d804faa7e449e8bd7df.png?version=1&modificationDate=1429930793740&api=v2 的定义，

http://wiki.baidu.com/download/attachments/86029397/411a4ffef87535ea11e59ec06c56b804.png?version=1&modificationDate=1429930793740&api=v2

http://wiki.baidu.com/download/attachments/86029397/dd7b2845fc67ee804d4faf3d7fe1f3e2.png?version=1&modificationDate=1429930793740&api=v2比较困难，因为我们不知道如何计算这种条件相关性（这是我们的最终目标）。 参考条件概率的计算方法，

http://wiki.baidu.com/download/attachments/86029397/f8ef311ce6f09230b1ca9f28f5c2de2c.png?version=1&modificationDate=1429930793741&api=v2

http://wiki.baidu.com/download/attachments/86029397/fdb27714e3d4d443134c6831de63ed20.png?version=1&modificationDate=1429930793741&api=v2比较抽象，其含义是 http://wiki.baidu.com/download/attachments/86029397/6133bba9db8422ee2585f012e0bfc095.png?version=1&modificationDate=1429930793741&api=v2的需求分散程度。即给定字符串的需求分布越分散，其与另一相关字符串的条件相关性越小，反则反之。例如，“医院”的需求分布相比“感冒”更分散，那么Rel(感冒|医院)应该小于Rel(医院|感冒)。非常类似BM25中DF的作用。

为什么不直接对 http://wiki.baidu.com/download/attachments/86029397/e316677b7b20de2f9ba69177c1758d47.png?version=1&modificationDate=1429930793741&api=v2按照条件概率的计算方法进行计算呢？因为我们不知道应该如何计算 http://wiki.baidu.com/download/attachments/86029397/c9de1b33c287cc7717aad54a4de0cdd4.png?version=1&modificationDate=1429930793741&api=v2。尽管按照其逻辑含义，可以设计很多方法，例如直接使用熵表示需求分散程度。但是，与前面一样，我们不知道对原有相关性的序关系会造成何种破坏，是否得不偿失。

那么， http://wiki.baidu.com/download/attachments/86029397/fdb27714e3d4d443134c6831de63ed20.png?version=1&modificationDate=1429930793741&api=v2怎么计算呢？这里有一个重要假设，“不管基于什么方法计算需求分散程度，基向量 http://wiki.baidu.com/download/attachments/86029397/df1445fbd51b0a95057b419167a368fe.png?version=1&modificationDate=1429930793742&api=v2的 http://wiki.baidu.com/download/attachments/86029397/fdb27714e3d4d443134c6831de63ed20.png?version=1&modificationDate=1429930793742&api=v2都是常数”（如果是0，可以设为一个很小的数近似处理）。常数不影响相关性计算的序关系，因此可以直接丢掉这一项。此时，

http://wiki.baidu.com/download/attachments/86029397/d2943957133699121cefa743ac50fbe3.png?version=1&modificationDate=1429930793742&api=v2

http://wiki.baidu.com/download/attachments/86029397/470f43c78251e64193360e0845c2003f.png?version=1&modificationDate=1429930793742&api=v2属于对称相关性，直接使用传统计算方式，例如cosine，那么

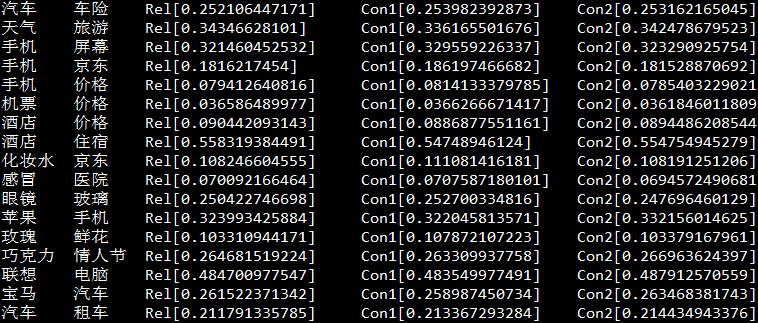
http://wiki.baidu.com/download/attachments/86029397/681e2df9ac90600aa4ce3bf332958cda.png?version=1&modificationDate=1429930793743&api=v2

式中 http://wiki.baidu.com/download/attachments/86029397/8057ed7fa53e640ddd534ff38fe218d1.png?version=1&modificationDate=1429930793743&api=v2的绝对值符号消失了，是因为基向量中的 http://wiki.baidu.com/download/attachments/86029397/58a26ae8459c7c132192081b29684666.png?version=1&modificationDate=1429930793743&api=v2与 http://wiki.baidu.com/download/attachments/86029397/8057ed7fa53e640ddd534ff38fe218d1.png?version=1&modificationDate=1429930793743&api=v2符号一致。

至此，得到了vector-based非对称条件相关性的计算方法，虽然理论不够严谨，但是尽量保证客观合理。直观上，假设d，q都进行了L2范数归一化，新方法相比cosine额外除了一项 http://wiki.baidu.com/download/attachments/86029397/89233332910896b239fd4013ab6a62c2.png?version=1&modificationDate=1429930793743&api=v2。进行L2范数归一化后，分布越分散，L1范数越大，符合预期。

**评估**

选取了一些term pairs，分别计算Rel(t1,t2), Rel(t2|t1), Rel(t1|t2)进行对比，结果如下（只关注Rel(t2|t1)与Rel(t1|t2)的相对大小即可，绝对值受word2vec和常数选择影响，不必关注）。使用的word2vec词表来自邵毅，在此感谢！



能够看到一些符合sense的结果，例如Rel(医院|感冒)>Rel(感冒|医院)，Rel(京东|手机)>Rel(手机|京东)，Rel(天气|旅游)>Rel(旅游|天气)等，也有一些说不清楚的case（酒店，价格等）。感兴趣的同学可以一起进一步讨论研究，也可以看看在已有应用中，非对称的相关性计算是否能够提升效果。

调研中暴露出的另一个问题是，很多现有的相关性算法对于事物之间的联系仍然难以表示，这也是后续值得深入研究的地方。