Modifier Graph Based Subtopic Mining

Haitao Yu Fuji Ren
yu-haitao@iss.tokushima-u.ac.jp ren@is.tokushima-u.ac.jp
Faculty of Engineering, University of Tokushima, 2-1 Minamijyousanjima-cho, Tokushima, Japan

❖ Intent Role Oriented Query Parsing

Two intent roles proposed by Yu et al. [1] are defined as:
Kernel-object (ko) refers to the dominant word that abstracts the core object of the underlying topic encoded within a query.
Modifier (mo) refers to the co-appearing words with kernel-object, which explicitly specify user's interested attributes or concrete aspects.
A query that can be represented with kernel-object and modifier is defined as **role-explicit**. Otherwise, it defined as **role-implicit**.

For subtopic mining, a subtopic string (denoted as $tStr$) can either be a real query or a query-like string obtained from other resources. Hence we can perform intent role annotation analogously.

Table 1. Intent role annotation for subtopic strings

Subtopic String	Intent Role Annotation		SogouQ
Harry Potter game	ko: Harry Potter	mo: game	In SogouQ
Harry Potter fiction	ko: Harry Potter	mo: fiction	In SogouQ
Harry Potter reading	ko: Harry Potter	mo: reading	Not in SogouQ

Definition 1 (**Co-Kernel-object Elements**) For one specific kernel-object, there exists a set of role-explicit subtopic strings that share the same kernel-object. We say these subtopic strings are co-kernel-object elements, denoted as $CoKO(ko) = \{tStr\}$. The elements in $CoKO(ko)$ are viewed as a set of expressions of the same kernel-object oriented subtopics.

Definition 2 (**Modifier Graph**) Modifier graph is an undirected, weighted graph $G_{mo} = (V, E, f)$ derived from a set of co-kernel-object elements $CoKO(ko)$, where: (i) The set of nodes is $V = \{mo\}$, namely the distinct modifiers in $CoKO(ko)$; (ii) $E = \{e | e = (mo_i, mo_j)\}$ is the set of undirected edges; (iii) $f(mo_i, mo_j, u) \rightarrow R$ is a function that assigns a weight. The parameter u is the set of scenarios that the function works.

Definition 3 (**Word-Level Co-Session**) For two distinct words w_i, w_j, if $\exists tStr_m \in Q$, $\exists tStr_n \in Q$ that meet $CoSessoin(tStr_m, tStr_n) \wedge w_i \in tStr_m \wedge w_j \in tStr_n$, we say w_i, w_j are co-session words.

Definition 4 (**Word-Level Co-Click**) For two distinct words w_i, w_j, if $\exists tStr_m \in Q$, $\exists tStr_n \in Q$ that meet $CoClick(tStr_m, tStr_n) \wedge w_i \in tStr_m \wedge w_j \in tStr_n$, we say w_i, w_j are co-click words.

Definition 5 (**Co-Parent**) For two distinct words w_i, w_j, if $\exists tStr$ that meets $w_i \in tStr$, $w_j \in tStr$, we say words w_i, w_j are co-parent words.

❖ Modifier Graph Clustering

Key Idea: Modifier graph is decomposable into clusters with strong intra-cluster interaction and relatively weak inter-cluster interaction. Each modifier cluster reasonably represents a possible subtopic.

The modifier associations among **游**戏 (game), 小说 (fiction) and 阅读 (reading) based on SogouQ are computed as:

Table 2. Modifier associations (co-parent : co-session : co-click)

Modifier matrix	**游**戏(game)	小说(fiction)	阅读(reading)
游戏 (game)	×	7:78:20	1:13:1
小说 (fiction)	7:78:20	×	700:993:3637
阅读 (reading)	1:13:1	700:993:3637	×

The corresponding modifier graph is:

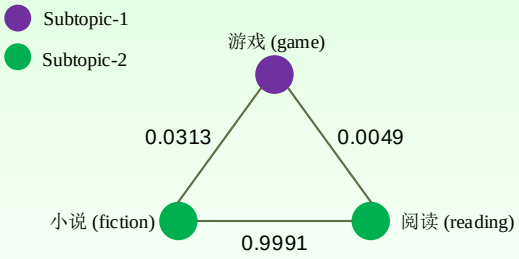

Figure 1. Modifier graph clustering

小说 (fiction) and 阅读 (reading) are strongly interacted (with a value of 0.9991). **游**戏 (game) and 小说 (fiction), **游**戏 (game) and 阅读 (reading) are weakly interacted. When performing graph clustering, we tend to group 小说 (fiction) and 阅读 (reading) into one cluster, and group **游**戏 (game) into another cluster. Due to the same kernel-object, the modifiers 阅读 (reading), **游**戏 (game) and 小说 (fiction) are positive indicators reflecting **哈利波特** (Harry Potter) oriented subtopics. Moreover, the co-session and co-click information are commonly interpreted as wisdom of crowds about search result preference and relevance [2-3]. It is reasonable to deduce that the two clusters indicate different subtopics, as well as the subtopics expressed by their parent subtopic strings.