# Text representation with Higher-Order Paths

Advantages of using higher-order paths [1] between documents are illustrated in Fig. 1. In this figure, there are three documents, *d1*, *d2* and *d3* which include a set of terms *{t1, t2, t3}, {t3, t4, t5}* and *{t4, t5}* respectively. Using a traditional similarity measure which is based on the shared terms (e.g. dot product), similarity value between documents *d1* and *d3* will be zero since they do not share any terms. But in fact these two documents have some similarities in the context of the dataset through *d2* as it can be seen in Fig. 1. This supports the idea that using higher-order paths between documents, it is possible to obtain a non-zero similarity value between *d1* and *d3* which was not possible in traditional Bag of Words (BOW) [2] representation. This value becomes larger if there are many interconnecting documents like *d2* between *d1* and *d3*. This may stem from the reason that the two documents are written on the same topic using two different but semantically closer sets of terms.

This project aims to represent these higher-order paths by using ***Linked-Lists***. Consequently, this project is a programming assignment in C, which aims to build an algorithm based on linked-lists that will build an efficient representation of documents.

***d1*** *{t1, t2}*

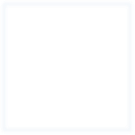
***d2***

***d3*** *{t4, t5}*

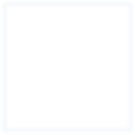
*{t2, t3, t4}*

**Fig. 1.** a) Illustration of higher-order paths

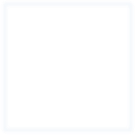
d1 d2 d3



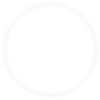
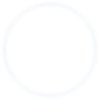
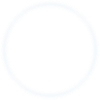
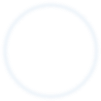
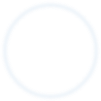
t1 t2



t2 t3 t4



t4 t5



2nd-order

2nd-order

d1

d2

d2

d3

t1 t2 t3

t4

t5

1st -order 1st -order 1st -order 1st -order

2nd-order

2nd-order 3rd-order

1st-order term co-occurrence {t1, t2}, {t2, t3}, {t3, t4}, {t2, t4}, {t4, t5} 2nd-order term co-occurrence {t1, t3}, {t1, t4}, {t2, t5}, {t3, t5}

3rd-order term co-occurrence {t1, t5}

**Fig. 2.** b) Graphical demonstration of first-order, second-order and third-order paths between terms through documents

First order :

{ilk,destek},{ilk,arka},{ilk,alarak},{ilk,orman},{ilk,sendika},{ilk,bilim},{ilk,adam},{ilk,daha},{ilk,insan},{ilk,embriyo},{ilk,klon},{ilk,ssk},{ilk,ilaç},{ilk,satış},{ilk,ilgi},{ilk,yazı},{destek,arka},{destek,alarak},{destek,orman},{destek,sendika},{destek,ankara},{destek,uluslararası},{arka,alarak},{arka,orman},{arka,sendika},{alarak,orman},{alarak,sendika},{orman,sendika},{ankara,uluslararası},{ankara,geç},{ankara,ay},{ankara,avrupa},{ankara,futbol},{ankara,şampiyona},{geç,ay},{geç,avrupa},{geç,futbol},{geç,şampiyona},{ay,avrupa},{ay,futbol},{ay,şampiyona},{avrupa,futbol},{avrupa,şampiyona},{futbol,şampiyona},{bilim,adam},{bilim,daha},{bilim,insan},{bilim,embriyo},{bilim,klon},{bilim,ssk},{adam,daha},{adam,insan},{adam,embriyo},{adam,klon},{adam,ssk},{daha,insan},{daha,embriyo},{daha,klon},{daha,ssk},{insan,embriyo},{insan,klon},{insan,ssk},{embriyo,klon},{embriyo,ssk},{klon,ssk},{ssk,ilaç},{ssk,satış},{ssk,ilgi},{ssk,yazı},{ilaç,satış},{ilaç,ilgi},{ilaç,yazı},{satış,ilgi},{satış,yazı},{ilgi,yazı},{ilgi,özcan},{ilgi,deniz},{ilgi,ve},{ilgi,naz},{ilgi,elmas},{ilgi,başrol},{ilgi,dizi},{ilgi,çekim},{özcan,deniz},{özcan,ve},{özcan,naz},{özcan,elmas},{özcan,başrol},{özcan,dizi},{özcan,çekim},{deniz,ve},{deniz,naz},{deniz,elmas},{deniz,başrol},{deniz,dizi},{deniz,çekim},{ve,naz},{ve,elmas},{ve,başrol},{ve,dizi},{ve,çekim},{naz,elmas},{naz,başrol},{naz,dizi},{naz,çekim},{elmas,başrol},{elmas,dizi},{elmas,çekim},{başrol,dizi},{başrol,çekim},{dizi,çekim}

Second order :

{ilk,ankara},{ilk,uluslararası},{ilk,özcan},{ilk,deniz},{ilk,ve},{ilk,naz},{ilk,elmas},{ilk,başrol},{ilk,dizi},{ilk,çekim},{destek,bilim},{destek,adam},{destek,daha},{destek,insan},{destek,embriyo},{destek,klon},{destek,ssk},{destek,ilaç},{destek,satış},{destek,ilgi},{destek,yazı},{destek,geç},{destek,ay},{destek,avrupa},{destek,futbol},{destek,şampiyona},{arka,bilim},{arka,adam},{arka,daha},{arka,insan},{arka,embriyo},{arka,klon},{arka,ssk},{arka,ilaç},{arka,satış},{arka,ilgi},{arka,yazı},{arka,ankara},{arka,uluslararası},{alarak,bilim},{alarak,adam},{alarak,daha},{alarak,insan},{alarak,embriyo},{alarak,klon},{alarak,ssk},{alarak,ilaç},{alarak,satış},{alarak,ilgi},{alarak,yazı},{alarak,ankara},{alarak,uluslararası},{orman,bilim},{orman,adam},{orman,daha},{orman,insan},{orman,embriyo},{orman,klon},{orman,ssk},{orman,ilaç},{orman,satış},{orman,ilgi},{orman,yazı},{orman,ankara},{orman,uluslararası},{sendika,bilim},{sendika,adam},{sendika,daha},{sendika,insan},{sendika,embriyo},{sendika,klon},{sendika,ssk},{sendika,ilaç},{sendika,satış},{sendika,ilgi},{sendika,yazı},{sendika,ankara},{sendika,uluslararası},{uluslararası,geç},{uluslararası,ay},{uluslararası,avrupa},{uluslararası,futbol},{uluslararası,şampiyona},{bilim,ilaç},{bilim,satış},{bilim,ilgi},{bilim,yazı},{adam,ilaç},{adam,satış},{adam,ilgi},{adam,yazı},{daha,ilaç},{daha,satış},{daha,ilgi},{daha,yazı},{insan,ilaç},{insan,satış},{insan,ilgi},{insan,yazı},{embriyo,ilaç},{embriyo,satış},{embriyo,ilgi},{embriyo,yazı},{klon,ilaç},{klon,satış},{klon,ilgi},{klon,yazı},{ssk,özcan},{ssk,deniz},{ssk,ve},{ssk,naz},{ssk,elmas},{ssk,başrol},{ssk,dizi},{ssk,çekim},{ilaç,özcan},{ilaç,deniz},{ilaç,ve},{ilaç,naz},{ilaç,elmas},{ilaç,başrol},{ilaç,dizi},{ilaç,çekim},{satış,özcan},{satış,deniz},{satış,ve},{satış,naz},{satış,elmas},{satış,başrol},{satış,dizi},{satış,çekim},{yazı,özcan},{yazı,deniz},{yazı,ve},{yazı,naz},{yazı,elmas},{yazı,başrol},{yazı,dizi},{yazı,çekim}

Third order :

{ilk,geç},{ilk,ay},{ilk,avrupa},{ilk,futbol},{ilk,şampiyona},{destek,özcan},{destek,deniz},{destek,ve},{destek,naz},{destek,elmas},{destek,başrol},{destek,dizi},{destek,çekim},{arka,özcan},{arka,deniz},{arka,ve},{arka,naz},{arka,elmas},{arka,başrol},{arka,dizi},{arka,çekim},{arka,geç},{arka,ay},{arka,avrupa},{arka,futbol},{arka,şampiyona},{alarak,özcan},{alarak,deniz},{alarak,ve},{alarak,naz},{alarak,elmas},{alarak,başrol},{alarak,dizi},{alarak,çekim},{alarak,geç},{alarak,ay},{alarak,avrupa},{alarak,futbol},{alarak,şampiyona},{orman,özcan},{orman,deniz},{orman,ve},{orman,naz},{orman,elmas},{orman,başrol},{orman,dizi},{orman,çekim},{orman,geç},{orman,ay},{orman,avrupa},{orman,futbol},{orman,şampiyona},{sendika,özcan},{sendika,deniz},{sendika,ve},{sendika,naz},{sendika,elmas},{sendika,başrol},{sendika,dizi},{sendika,çekim},{sendika,geç},{sendika,ay},{sendika,avrupa},{sendika,futbol},{sendika,şampiyona}, {ankara,bilim},{ankara,adam},{ankara,daha},{ankara,insan},{ankara,embriyo},{ankara,klon},{ankara,ssk},{ankara,ilaç},{ankara,satış},{ankara,ilgi},{ankara,yazı},{uluslararası,bilim},{uluslararası,adam},{uluslararası,daha},{uluslararası,insan},{uluslararası,embriyo},{uluslararası,klon},{uluslararası,ssk},{uluslararası,ilaç},{uluslararası,satış},{uluslararası,ilgi},{uluslararası,yazı},{bilim,özcan},{bilim,deniz},{bilim,ve},{bilim,naz},{bilim,elmas},{bilim,başrol},{bilim,dizi},{bilim,çekim},{adam,özcan},{adam,deniz},{adam,ve},{adam,naz},{adam,elmas},{adam,başrol},{adam,dizi},{adam,çekim},{daha,özcan},{daha,deniz},{daha,ve},{daha,naz},{daha,elmas},{daha,başrol},{daha,dizi},{daha,çekim},{insan,özcan},{insan,deniz},{insan,ve},{insan,naz},{insan,elmas},{insan,başrol},{insan,dizi},{insan,çekim},{embriyo,özcan},{embriyo,deniz},{embriyo,ve},{embriyo,naz},{embriyo,elmas},{embriyo,başrol},{embriyo,dizi},{embriyo,çekim},{klon,özcan},{klon,deniz},{klon,ve},{klon,naz},{klon,elmas},{klon,başrol},{klon,dizi},{klon,çekim}

TF VALUE :

Econ Health Magazin

ankara;2 ssk;2 çekim;1

destek;2 ilk;2 dizi;1

şampiyona;1 yazı;1 başrol;1

futbol;1 ilgi;1 elmas;1

avrupa;1 satış;1 naz;1

IDF VALUE :

Econ Health Magazin

şampiyona; 1,098612 ilgi; 0,693147 çekim; 0

avrupa; 1,098612 yazı; 0,693147 dizi; 0

futbol; 1,098612 satış; 0,693147 başrol; 0

ankara; 0,810930 ilk; 0 elmas; 0

destek; 0,810930 ssk; 0 naz; 0

BIG O :

If we say m ---> for master linked list’s size x ---> for second order linked list’s size

a ---> for first order linked list’s size y ---> for third order linked list’s size

the biggest loops are in the second\_order insertion and third\_order insertion

for second\_order insertion --- O(m\*a^2\*n) n = max(a,x)

for third\_order insertion --- O(m\*x\*a\*n) n = max(a,x,y)

if we say k = max(a,x)

---> O(m\*a\*n\*k)

Muhammed Enes AKTÜRK