RESEARCH ARTICLE

ການຈັດກຸ່ມເວບວິດີໂອຕາມເຕັກນິກຂອງການຮວມກຸ່ມດ້ວຍການນຳໃຊ້ຂໍ້ມູນ ປະເພດຂໍ້ຄວາມ (Text Information)*

ວິນັດ ເມກທະນາວັນ†

ພາກວິຊາວິສະວະກຳຄອມພິວເຕີ, ຄະນະວິສະວະກາສາດ, ມະຫາວິທະຍາໄລສຸພານຸວົງ, ສປປ ລາວ

ບົດຄັດຫຍໍ້

ຢູ່ໃນສື່ສັງຄົມອອນລາຍປະຈຸບັນ, ການຈັດກຸ່ມເວບວິດີໂອ ແມ່ນມີຄວາມທ້າທາຍຕົ້ນຕໍ ໃນການຈັດລຽງ ແລະ ການຕອບສະໜອງ ຄວາມເພິ່ງພໍໃຈ ຂອງຜູ້ຊົມໃຊ້ໃນການຮຽກເອີ້ນເບິ່ງວິດີໂອຢູ່ໃນເວບ. ດ້ວຍເຫດ ຜົນດັ່ງກ່າວ ບົດວິໄຈນີ້ຈຶ່ງໄດ້ນຳສະເໜີພື້ນຖານແນວຄວາມຄິດໃໝ່ ສຳລັບການຈັດກຸ່ມເວບວິດີໂອ ຢູ່ໃນສື່ສັງຄົມ ດ້ວຍການນຳໃຊ້ຄຸນລັກສະນະຂໍ້ມູນປະເພດຂໍ້ຄວາມ (textual features) ເຊັ່ນ: ຫົວຂໍ້ (title), ເນື້ອໃນ (tag), ຄາອະທິບາຍ (description) ແລະ ແຫຼ່ງສະໜັບສະໜູນ ຈາກພາຍນອກໂດຍອີງຕາມຄວາມສຳພັນ ກັນທາງດ້ານຄວາມໝາຍຂອງຄຳສັບ. ກ່ອນອື່ນໝົດ ພວກເຮົາໄດ້ປັບປຸງຟັງຊັນ ການຄິດໄລ່ຄວາມຄ້າຍຄືກັນ ລະຫວ່າງເອກະສານວິດີໂອແບບດັ່ງເດີມ ດ້ວຍການປັບລະດັບຄວາມສຳຄັນຂອງ (feature vectors) ຢູ່ໃນ Vector Space Model (VSM). ຫຼັງຈາກນັ້ນ ໄດ້ພິຈາລະນາຄວາມໝາຍຂອງຂໍ້ມູນ ລະຫວ່າງຄຳສັບຕ່າງໆ ໂດຍການໃຊ້ Word2vec ເພື່ອເກັບກຳການຕໍ່ເນື່ອງ ຂອງເວັກເຕີຂອງຄຳສັບຢູ່ໃນເອກະສານວິດີໂອ. ພ້ອມກັນນັ້ນ ກໍ່ໄດ້ກຳນົດຟັງຊັນໃໝ່ຂຶ້ນມາ ສຳລັບການຄິດໄລ່ຄວາມຄ້າຍຄືກັນ ລະ ຫວ່າງເອກະສານສອງເອກະ ສານ ໂດຍອີງໃສ່ເຕັກນິກ ຄວາມສາມາດໃນການສຳພັນກັນຕາມຄວາມໝາຍ ຂອງຄຳສັບທີ່ກ່ຽວຂ້ອງ ທີ່ນອນຢູ່ໃນແຕ່ລະໝວດຂອງເອກະສານ. ພາຍຫຼັງທີ່ໄດ້ຄ່າຄວາມຄ້າຍຄືກັນ ຂອງເອກະສານ ໃນແຕ່ລະໂມເດວແລ້ວ ພວກເຮົາໄດ້ກຳນົດຟັງຊັນ ການຮວມຄ່າຄວາມຄ້າຍຄືກັນເຫຼົ່ານີ້ ເຂົ້າກັນ ກ່ອນນຳໄປໃຊ້ໃນໂມເດວຂອງການຈັດກຸ່ມ. ສຸດທ້າຍ, ການປະຕິບັດການທົດລອງພວກເຮົາໃຊ້ເຕັກນິກ Clustering Ensemble ເພື່ອໃຫ້ໄດ້ເລເບວການຮວມກັນ ຂອງແຕ່ລະກຸ່ມ. ສຳລັບການປະເມີນການທົດລອງ ແມ່ນໄດ້ປະຕິບັດກັບຂໍ້ມູນຈິງ ໃນສື່ສັງຄົມ (YouTube) ເຊິ່ງຜົນຂອງການທົດລອງ ໄດ້ສະແດງໃຫ້ເຫັນເຖິງຂໍ້ດີ ຂອງແນວທາງການນຳສະເໜີໂມເດວ ການຈັດກຸ່ມຂອງເວບວິດີໂອ.

ຄຳສັບສຳຄັນ

ຄວາມສຳພັນດ້ານຄວາມໝາຍຂອງຂໍ້ມູນ, ຮວມການຈັດກຸ່ມ, ມາຍນິງເວບວິດີໂອ

---

* **ການອ້າງອີງພາສາລາວ**

ວິນັດ ເມກທະນາວັນ (2018). ການຈັດກຸ່ມ ເວບວິດີໂອ ຕາມເຕັກນິກ ຂອງການຮວມກຸ່ມ ດ້ວຍການນຳໃຊ້ຂໍ້ມູນ ປະເພດຂໍ້ຄວາມ (Text Information), ວາລະສານວິທະຍາສາດ ມະຫາວິທະຍາໄລສຸພານຸວົງ, ສະບັບທີ 3 : ໜ້າ 98-112.

† ***ຕິດຕໍ່ພົວພັນ***

ວິນັດ ເມກທະນາວັນ. ພາກວິຊາວິສະວະກຳຄອມພິວເຕີ, ຄະນະວິສະວະກາສາດ, ມະຫາວິທະຍາໄລສຸພານຸວົງ. Tel: 020 55407070, E-mail: vinath.mek@gmail.com

RESEARCH ARTICLE

Web Video Clustering based on Cluster Ensemble by using Text Information

Vinath Mekthanavanh‡

Department of Computer Engineering, Faculty of Engineering, Souphanouvong University, Lao PDR

ABSTRACT

In social media, web video categorization is the main challenge for organizing and retrieving the number of videos on the web. From this perspective, a novel framework for social web video clustering is proposed by using textual features, e.g., title, tag and description, and external support source based on semantic relation. Firstly, we modify the traditional similarity by adjusted weights of feature vectors in Vector Space Model (VSM). Secondly, by considering the semantic information between terms, the Word2vec is used to capture the continuous vectors of terms in a document. Thirdly, a new distance function for measuring the similarity between two documents is defined based on link-based Semantic Relation of the relevant terms in each category document. Fourthly, the combining function is approached to combine the similarities before applying them to clustering models. Finally, we employ the Clustering Ensemble to integrate the labels of each clustering. Experimental evaluations on real-world web video datasets demonstrate the advantages of the proposed approach.

Keywords

Text information, Clustering Ensemble, Web Videos Mining

‡ ***Correspondence***

Vinath Mekthanavanh, Department of Computer Engineering, Faculty of Engineering, Souphanouvong University, Tel: 020 55407070, E-mail: vinath.mek@gmail.com

1. ພາກສະເໜີ

ໃນສື່ສັງຄົມອອນລາຍ ປະຈຸບັນ ຢູທູບ (YouTube) ໄດ້ກາຍເປັນເວບໄຊ ທີ່ໄດ້ຮັບຄວາມນິຍົມຫຼາຍ ໃນອິນເຕີເນັດ ເຊິ່ງເປັນເວບໄຊ ທີ່ອະນຸຍາດ ໃຫ້ຜູ້ຊົມໃຊ້ ສາມາດອັບ�โหຼດວິດີໂອ ທີ່ຕົນເອງຕ້ອງການແຊ ໃຫ້ຜູ້ອື່ນໄດ້ຮັບຮູ້ ໄດ້ເຫັນນາໂດຍບໍ່ໄດ້ກຳນົດລະດັບ ຫຼື ວຸດທິ ຂອງຜູ້ທີ່ອັບໂຫຼດ ຈຶ່ງເປັນສາຍເຫດເຮັດໃຫ້ຈຳນວນ ຂອງວິດີໂອເພີ່ມຂຶ້ນຢ່າງໄວວາ ຕົວຢ່າງ: ພາຍຫຼັງ 3 ປີນັບຈາກຢູທູບ ໄດ້ອອກສູ່ສັງຄົມ ໃນເດືອນກຸມພາ 2005 ເຊິ່ງໃນເວລານັ້ນຈຳນວນ ຂອງວິດີໂອແມ່ນມີຫຼາຍກວ່າ 45 ລ້ານວິດີໂອທີ່ບັນຈຸ ໃນຄັງເກັບຂໍ້ມູນ, ມີການອັບ�โหຼດວິດີໂອ ທີ່ສັງລວມໄດ້ປະມານ 7 ຊົ່ວໂມງໃນທຸກໆນາທີ. ວິດີໂອໄດ້ເພີ່ມຂຶ້ນຢ່າງຕໍ່ເນື່ອງດ້ວຍ ຈຳນວນມະຫາສານ ໃນເດືອນມີຖຸນາ 2017 ຢູທູບ ໄດ້ມີຜູ້ເຂົ້າໃຊ້ ໃນແຕ່ລະເດືອນຫຼາຍກວ່າ 1.5 ພັນລ້ານຄົນ ຜູ້ຊົມໃຊ້ຈາກທົ່ວໂລກໄດ້ໃຊ້ເວລາ ໃນການເຂົ້າເບິ່ງວິດີໂອ ປະມານ ພັນລ້ານຊົ່ວໂມງ ແລະ ພວກເຂົາໄດ້ອັບ�โหຼດວິດີໂອ ຫຼາຍກວ່າ 500 ຊົ່ວໂມງໃນທຸກໆນາທີ. ການຈັດເກັບວິດີໂອຢູ່ໃນຢູທູບ ຈາກການອັບ�โหຼດ ໄດ້ເພີ່ມຂຶ້ນດ້ວຍອັດຕາສ່ວນທີ່ໜ້າຕົກໃຈປະມານ 8 ຊົ່ວໂມງ/ວິນາທີ ແລະ ການຈັດເກັບຈຳນວນຂອງວິດີໂອ ຢ່າງມະຫາສານນີ້ ໄດ້ກາຍເປັນບັນຫາທີ່ ສຳຄັນໃຫ້ແກ່ຜູ້ຊົມໃຊ້ ເພື່ອຄົ້ນຫາວິດີໂອທີ່ສອດຄ່ອງ ແລະ ຖືກຕ້ອງ ຕາມຄວາມຕ້ອງການ ຂອງພວກເຂົາ.

ສະນັ້ນ, ການແບ່ງກຸ່ມ ຂອງເວບວິດີໂອ ແບບອັດຕະ ໂນມັດແມ່ນປັດໄຈສຳຄັນ ໃນການ ລະບຸປະເພດ, ດັດສະນີ ແລະ ການຄົ້ນຫາວິດີໂອ ຢູ່ເທິງເວບ. ການຈັດກຸ່ມ ຫຼື ແບ່ງກຸ່ມ ຂອງວິດີໂອ ທີ່ມີຂໍ້ມູນຂະໜາດໃຫຍ່ ແລະ ວິດີໂອທີ່ມີຄວາມຫຼາກຫຼາຍ ແມ່ນເປັນສິ່ງທີ່ມີຄວາມທ້າທາຍຫຼາຍ. ໃນໄລຍະຜ່ານມາ ໄດ້ມີນັກຄົ້ນຄວ້າວິໄຈ ຈຳນວນຫຼວງຫຼາຍ ໄດ້ລົງເລິກໃນການວິເຄາະຂໍ້ມູນ ວິດີໂອ ເຊິ່ງແຕ່ລະບົດວາລະສານ ໃນຂົງເຂດການວິໄຈນີ້ ແມ່ນໄດ້ນຳໃຊ້ ຄຸນລັກສະນະ ຂອງຂໍ້ມູນທີ່ແຕກຕ່າງກັນເຊັ່ນ: ຂໍ້ຄວາມ, ຄວາມໝາຍຂອງຄາສັບ, ສຽງ (Audio) ແລະ ຮູບພາບ (Visual) ທັງໃຊ້ໃນ ຄຸນລັກສະນະດ່ຽວ, ປະສົມປະສານກັນຕາມ ຄວາມເໝາະສົມ ແລະ ຄວາມສາມາດເປັນໄປໄດ້ ເພື່ອເຮັດທົດລອງ ໃນໂມເດວ ຂອງການແບ່ງກຸ່ມ ເວບວິດີໂອ (Yang et al. 2007; Jiang et al. 2007) ແລະ ຫຼັງຈາກນັ້ນໄດ້ມີການພັດທະນາເພີ່ມ ເຕີມໂດຍ (Wang et al. 2010; Song et al. 2010) ດ້ວຍການນຳໃຊ້ ວິທີການແບ່ງກຸ່ມ ຈາກ ແຫຼ່ງຂໍ້ມູນທາງເວບ ແບບເປີດເພື່ອແບ່ງກຸ່ມ ວິດີໂອ ທີ່ມີຂະໜາດໃຫຍ່. ການແບ່ງກຸ່ມວິດີໂອ ໄດ້ຮັບ ການປັບປຸງໂດຍ (Toderici et at. 2010), (Schindler et al. 2008) ໄດ້ຄົ້ນຄິດວິທີ ການ ແທນຂໍ້ມູນວິດີໂອ ດ້ວຍຊຸດຄຳສັບ ທີ່ມີການຮວມ ກັນ ຂອງຄຳອະທິບາຍ ສະພາບແວດລ້ອມ ຕົວຈິງ ແລະ ເວລາ. ນອກຈາກນັ້ນ ການຈັດໝວດວິດີໂອ ແບບ ອັດຕະໂນມັດ ແມ່ນໄດ້ນຳໃຊ້ເຕັກນິກ ຂອງ SVM ຮ່ວມກັບການໃຊ້ຄຸນລັກສະນະ ລະດັບຕ່ຳ ຂອງວິດີໂອຄື (low-level features) ເພື່ອດຳເນີນ ການແບ່ງກຸ່ມ (Yuan et al. 2006), (Wu et al. 2009) ນຳໃຊ້ຂໍ້ມູນຕາມບໍລິບົດ (Contextual) ແລະ ຂໍ້ມູນຕາມສື່ສັງຄົມທີ່ອັບ�โหຼດ ໂດຍຜູ້ຊົມໃຊ້ ແລະ ນຳໃຊ້ຂໍ້ມູນວິດີໂອ ທີ່ກ່ຽວຂ້ອງເຊິ່ງໄດ້ລະບຸ ໄວ້ໂດຍຢູທູບ ສຳລັບການຈັດປະເພດ ຂອງເວບວິ ດີໂອ. ປະສິດທິພາບ ໃນການຈັດປະເພດວິດີໂອ ໄດ້ມີການປັບປຸງເພີ່ມເຕີມ ດ້ວຍການສັງລວມ ຄວາມໝາຍຕາມບໍລິບົດ ແລະ ຄວາມໝາຍ ຂອງວິ ດີໂອ, ວິດີໂອທີ່ກ່ຽວຂ້ອງ ແລະ ຄວາມສົນໃຈ ຂອງ ຜູ້ໃຊ້ຕາມການຄາດຄະເນ ໃນການຈັດວິດີໂອ ໃຫ້ ຖືກຕ້ອງຕາມໝວດ ຂອງເວບວິດີໂອ (Wu et al. 2012). (Chen et al. 2010) ໃຊ້ເນື້ອໃນແບບ ເປີດທີ່ມີການຊ້ຳຊ້ອນກັນ ແລະ ໝວດຂອງວິກິພີ ເດຍ (Wikipedia) ສຳລັບການຈັດປະເພດ ຂອງ ເວບວິດີໂອ, (Mahmood et al. 2015) ໄດ້ນຳ ສະເໜີໂມເດວ ການຈັດປະເພດເວບວິດີໂອ ໂດຍ ອີງຕາມພື້ນຖານ ການຈັດກຸ່ມ ດ້ວຍການນຳໃຊ້ ຂໍ້

ມູນ ປະເພດຂໍ້ຄວາມ ແລະ ແຫຼ່ງສະໜັບສະໜູນ ຈາກພາຍນອກ (Google). ຈາກນັ້ນ ບົນພື້ນຖານ ໃນການແກ້ໄຂບັນຫາ ທີ່ມີເປົ້າໝາຍດຽວກັນ ກໍ່ຄື ການຈັດກຸ່ມວີດີໂອ ໂດຍແນໃສ່ການເພີ່ມ ປະສິດ ທິພາບຂອງການທົດລອງ ໃນໂມເດວຕົວຈິງ, ໂມ ເດວການຈັດກຸ່ມເວບວີດີໂອ ໄດ້ຖືກນຳສະເໜີຂຶ້ນ ມາ (Mekthanavanh and Li 2016) ໂດຍການ ໃຊ້ຄຸນລັກສະນະ ຂອງຂໍ້ມູນທີ່ແຕກຕ່າງກັນ ຕົວ ຢ່າງ: ຂໍ້ຄວາມ, ຄວາມໝາຍຂອງຄຳສັບ ແລະ ຮູບ ພາບ ເຊິ່ງຖືຢູ່ໃນຮູບແບບຂອງ multi-modality.

ພາຍໃຕ້ການນຳສະເໜີ ການວິໄຈ ໃນ ສາຂາ ຂອງ text mining, ຢູ່ໃນໂມເດວດັ່ງກ່າວນີ້ ຄວາມຄ້າຍຄືກັນ ລະຫວ່າງຄຸນລັກສະນະ ຂອງ ສອງ textual ເວັກເຕີຖືວ່າເປັນປັດໄຈສຳຄັນ. ຂໍ້ ມູນປະເພດ ຂໍ້ຄວາມ ເຊິ່ງຜູ້ໃຊ້ໄດ້ອັບໂຫຼດຂຶ້ນ ພ້ອມກັບວີດີໂອນັ້ນ ເຊິ່ງປະກອບມີ: title, tag ແລະ description ແມ່ນລວມມີທັງຄຳສັບທີ່ມີ ຄວາມໝາຍ, ບໍ່ມີຄວາມໝາຍ, ຄາເວົ້າມີການຄຸມ ເຄືອກັນບໍ່ຈະແຈ້ງ ຫຼື ບາງເທື່ອກໍ່ຂາດຫາຍໄປ ເນື່ອງຈາກວ່າ ການຂຽນຄຳອະທິບາຍ ໃຫ້ກັບວີດີ ໂອ ໃນເວລາອັບໂຫຼດ ແຕ່ລະເທື່ອນັ້ນ ແມ່ນຂຶ້ນ ກັບຜູ້ໃຊ້. ຕໍ່ກັບບັນຫາດັ່ງກ່າວ ເຮັດໃຫ້ຂໍ້ມູນທີ່ຈະ ນຳໄປວິໄຈ ມີປະສິດທິຜົນທີ່ຕ່ຳ ແລະ ຜົນ ຂອງ ການວິໄຈ ກໍ່ບໍ່ດີເທົ່າທີ່ຄວນ. ສະນັ້ນ, ເພື່ອເຮັດ ໃຫ້ຂໍ້ມູນດັ່ງກ່າວ ມີປະສິດທິຜົນທີ່ສູງຂຶ້ນ ແລະ ສິ່ງຜົນທີ່ດີຕໍ່ກັບ ຜົນໄດ້ຮັບນັ້ນ ມັນຈຳເປັນຕ້ອງໄດ້ ຮັບການສະໜັບສະໜູນ ຈາກແຫຼ່ງຂໍ້ມູນ ພາຍ ນອກ ໂດຍໃຫ້ເຊື່ອມຍົງ ກັບຂໍ້ມູນເດີມ ທີ່ມີຢູ່ ເພື່ອເປັນການເຊື່ອມຄວາມສຳພັນກັນ ທາງດ້ານ ຄວາມໝາຍ ຂອງຄຳສັບ ລະຫວ່າງຄາສັບດ້ວຍກັນ ພາຍໃນ 1 ຊຸດຂໍ້ມູນໃຫ້ມີຄວາມໝາຍຂຶ້ນຕື່ມ. ໂດຍອີງໃສ່ພື້ນຖານ ຄວາມສຳພັນກັນ ລະຫວ່າງຄຳ ສັບເຫຼົ່ານັ້ນ (Cilibrasi and Vitanyi 2006) ໄດ້ ພັດທະນາ ວິທີການກວດສອບ ຄວາມສຳພັນກັນ ລະຫວ່າງສອງຄຳສັບ ຫຼື ສອງປະໂຫຍກ ດ້ວຍການ ໃຊ້ວິທີການຄົ້ນຫາ ຂອງ Google ແລະ ພວກເຂົາ ໄດ້ສືບຕໍ່ພັດທະນາຕົວບົ່ງຊີ້ທາງສະຖິຕິຂຶ້ນອີກ ໂດຍອີງຕາມຈຳນວນທີ່ສົ່ງຜົນຕອບຮັບຄືນໃນ Google page count ເພື່ອຄິດໄລ່ຄວາມຄ້າຍຄື ກັນລະຫວ່າງສອງຄຳສັບໃນຮູບແບບຂອງ World Wide Web ເຊິ່ງເອີ້ນວ່າ Normalize Google Distance (NGD)(Cilibrasi and Vitanyi 2007). ພ້ອມດຽວກັນນີ້ NGD ຍັງໄດ້ຖືກນຳໃຊ້ ເຂົ້າໃນການຄິດໄລ່ຄວາມຄ້າຍຄືກັນຂອງຄວາມ ສຳພັນກັນຂອງຄຳສັບໃນ recommendation model (Huang et al. 2016).

ໃນປະຈຸບັນນີ້, ໂມເດວ word2vec ໄດ້ ສ້າງໂອ ກາດໃໝ່ໃຫ້ແກ່ການວິໄຈໃນສາຂາຂອງ text mining ເປັນໂຄງການ ການພັດທະນາຊອບ ແວທີ່ເປີດກວ້າງສູ່ສັງຄົມ ເຊິ່ງນຳສະເໜີໂດຍ (Mikolov et al. 2013). Word2vec ສາມາດ ຈັບຄູ່ກັນ ລະຫວ່າງຄຳສັບກັບເວັກເຕີຂອງມັນເອງ ໄດ້ຢ່າງຖືກຕ້ອງ ແລະ ຖືໄດ້ວ່າເປັນຄ່າປະມານທີ່ ສົມບູນສຳລັບການສະແດງຄຳສັບຢູ່ໃນຮູບແບບ ຂອງເວັກເຕີ ແລະ ສາມາດສະແດງຄວາມໝາຍ ພິເສດທີ່ເປັນປະໂຫຍດ ໃນການຈັດແບ່ງກຸ່ມ ຂໍ້ຄວາມ (text) (Lilleberg et al. 2015). ນອກຈາກນີ້ມັນຍັງຊ່ວຍໃນການປັບປຸງຄວາມ ຖືກຕ້ອງໃນການຈັດປະເພດຂອງເອກະສານ (Ju et al. 2015). ຢູ່ໃນບົດວິໄຈນີ້, ພວກເຮົາໃຊ້ word2vec (Skip-gram model) ເພື່ອສະແດງ ເອກະສານວີດີໂອໃນຮູບແບບຂອງເວັກເຕີດ້ວຍ ການຄູນເວັກເຕີທີ່ຢູ່ໃນແຕ່ລະເອກະສານກັບຄ່າ Term Frequency-Inverse Document Frequency (TF-IDF) ແລະ ຫຼັງຈາກນັ້ນຄິດໄລ່ ຄ່າສະເລ່ຍຂອງພວກມັນ.

ພ້ອມນີ້ ເພື່ອໃຫ້ໄດ້ຄວາມສາພັນກັນທາງ ດ້ານຄວາມໝາຍຂອງຄຳສັບພາຍໃນຄັງຂອງຂໍ້ມູນ, ໂມເດວໄດ້ແບ່ງອອກເປັນ 2 ກຸ່ມຄື: ໂມເດວ ຄວາມສຳພັນແບບການຈັບຄູ່ (pairwise relation model) ແລະ ຄວາມສາພັນແບບຄ້າຍຄື ກັ ນ (similarity-based relation model). ສຳລັບຄວາມສຳພັນແບບຈັບຄູ່, ຖ້າສອງຄາສັບ

ຫາກປະກົດຂຶ້ນມານຳກັນຫຼາຍໆຄັ້ງ ສະແດງວ່າ ພວກມັນມີຄວາມສຳພັນເຊິ່ງກັນ ແລະ ກັນ, ແລະ ສາມາດສ້າງຄວາມສຳພັນດ້ວຍຕົວພວກມັນເອງ ແຕ່ບໍ່ສາມາດສ້າງຄວາມສຳພັນອື່ນໆທີ່ນອກເໜືອ ຈາກນີ້ໄດ້. ຄວາມສຳພັນແບບຄ້າຍຄືກັນ, ຖ້າຫາກ ວ່າສອງຄຳສັບທີ່ມີຄວາມຄ້າຍຄືກັນ ເຊິ່ງມີຄວາມ ຖີ່ການປະກົດຂຶ້ນມາຄືກັບຄຳສັບອື່ນ ສະແດງວ່າ ສອງຄຳສັບນີ້ແມ່ນມີຄວາມໝາຍອັນດຽວກັນ. ດັ່ງນັ້ນ, ຢູ່ໃນຄວາມສາພັນແບບຈັບຄູ່ມັນຈາເປັນ ຕ້ອງມີການປະຕິສຳພັນກັບຄຳສັບອື່ນໆເພື່ອໃຫ້ມີ ຄວາມເຂັ້ມແຂງໃນການຄຳນວນຄວາມສຳພັນຂອງ ຄຳສັບໃນລັກສະນະນີ້. ຈາກປະສິດທິພາບ ຂອງ ໂມເດວຄວາມສຳພັນດ້ານຄວາມໝາຍນີ້ (Cheng et al. 2015) ໄດ້ນຳສະເໜີການວິເຄາະຄວາມ ສຳພັນແບບຈັບຄູ່ຂອງຂໍ້ມູນທີ່ເກີດຂຶ້ນ ເຊິ່ງໄດ້ ຈາກການສັງລວມເອກະສານ ແລະ ການວິເຄາະ ຄວາມສຳພັນແບບຄ້າຍຄືກັນດ້ວຍການປະຕິ ສຳພັນກັບຄຳສັບອື່ນໆພາຍໃນເອກະສານດຽວກັນ.

ການຮວມກຸ່ມ (Clustering ensemble) ເປັນວິທີການທົ່ວໄປທີ່ໃຊ້ເຂົ້າໃນການແກ້ໄຂ ບັນຫາການຈັດກຸ່ມ ເຊິ່ງມັນເປັນການລວມກຸ່ມ ຍ່ອຍເຂົ້າກັນ ໃຫ້ກາຍເປັນກຸ່ມທີ່ມີປະ ສິດທິພາບ ທີ່ດີກວ່າ ແລະ ວິທີການນີ້ຈະຊ່ວຍປັບປຸງຄວາມ ຖືກຕ້ອງ, ຄວາມສະຖຽນລະພາບ, ຄວາມແຂງ ແກ່ນ ແລະ ຄວາມໜ້າເຊື່ອຖືດ້ວຍການຄາດຄະເນ ການລວມຜົນໄດ້ຮັບຂອງການຈັດກຸ່ມຫຼາຍກຸ່ມ ໂດຍໃຊ້ອາກໍຣິທຶມດຽວກັນ ຕາມຄ່າຂອງພາຣາມິ ເຕີທີ່ແຕກຕ່າງກັນ ຫຼື ການລວມຜົນໄດ້ຮັບການ ຈັດກຸ່ມທີ່ໃຊ້ອາກໍຣິທຶມຕ່າງກັນ ດ້ວຍການໃຊ້ຊຸດ ຂໍ້ມູນດຽວກັນ (Topchy et al. 2005). ນອກຈາກນີ້ (Strehl and Ghosh 2002) ໄດ້ນຳ ສະເໜີຮູບແບບຂອງການຮວມກຸ່ມ ດ້ວຍການນຳ ໃຊ້ທິດສະດີຂອງກຣາບ (graph) ເພື່ອສັງລວມ ຜົນຮັບຂອງການຈັດກຸ່ມ, (Fred and Jain 2005) ໃຊ້ຫຼັກການ ການຮວມກຸ່ມຫຼາຍກຸ່ມເຂົ້າ ດ້ວຍກັນ ແລະ ສະ ແດງໃຫ້ເຫັນວ່າຜົນຂອງ ການຮວມກຸ່ມດັ່ງກ່າວແມ່ນມີປະສິດ ທິພາບດີກ ວ່າວິທີການຮວມກຸ່ມອື່ນໆ. ການຮວມກຸ່ມ ໄດ້ມີ ການນຳໃຊ້ຢ່າງກວ້າງຂວາງເຂົ້າໃນການຄົ້ນຄວ້າ ວິໄຈໃນຂອບເຂດຂອງ data mining ກ່ຽວກັບ ໂປຣແກຣມນຳໃຊ້ເຊັ່ນວ່າ web analysis, recommendation system, text mining, network security, pattern recognition ແລະ ອື່ນໆ.

ຈາກເຫດຜົນ ແລະ ຂໍ້ດີຈາກຜົນຂອງ ການເຮັດວິໄຈໃນຫຼາຍໆບົດທີ່ໄດ້ກ່າວມາຂ້າງເທິງ ນີ້, ມັນຈຶ່ງມີຄວາມຈຳ ເປັນໃນການພິຈາລະນາ ແລະ ມີການຄົ້ນຄວ້າຊອກຫາວິທີທາງເພື່ອປັບປຸງ ລະບົບການຄົ້ນຫາວິດີໂອ ຢູ່ໃນສື່ສັງຄົມຢູທູບໃຫ້ ມີຄວາມສາມາດຕອບສະໜອງຄວາມຕ້ອງການ ຂອງຜູ້ຊົມ ໃຊ້ໃຫ້ດີຂຶ້ນ ແລະ ຫຼຸດຜ່ອນຄວາມ ຜິດພາດໃນການຄົ້ນຫາໃຫ້ໜ້ອຍລົງ. ຢູ່ໃນບົດ ວິໄຈນີ້ ພວກເຮົາເຊື່ອໝັ້ນວ່າການປັບປຸງຄວາມ ສາມາດໃນການຈັດກຸ່ມ ດ້ວຍການປ່ຽນແປງສູດ ການຄິດໄລ່ຄວາມຄ້າຍຄືກັນຂອງເອກະສານວິດີ ໂອຢູ່ໃນ VSM, ການກຳນົດສູດໃໝ່ຂຶ້ນມາເພື່ອ ຄິດໄລ່ໄລຍະຄວາມສຳພັນກັນຂອງວິດີໂອໂດຍ ອີງໃສ່ຫຼັກການ ການວິເຄາະຄວາມສາພັນດ້ານ ຄວາມໝາຍຂອງ link-based ຢູ່ໃນ (Cheng et al. 2015) ເພື່ອສົມທຽບເອກະສານວິດີໂອຕາມ ແຕ່ລະໝວດຂອງເອກະສານ, ແລະ ການນາໃຊ້ word2vec ເພື່ອໃຫ້ໄດ້ຂໍ້ມູນທີ່ມີຄວາມໝາຍ ດ້ານເນື້ອໃນຢ່າງແທ້ຈິງຂອງຄຳສັບນັ້ນໆ ແມ່ນຈະ ສາມາດຊ່ວຍໃນການສະແດງຄວາມໝາຍຂອງວິດີ ໂອ ໄດ້ຢ່າງຈະແຈ້ງຂຶ້ນຕື່ມ ເຊິ່ງເຮັດໃຫ້ການໄຈ້ ແຍກ ແລະ ການຈັດກຸ່ມມີປະສິດທິພາບດີຂຶ້ນ.

2. ນິຍາມ ແລະ ວິທີການ

2.1. ນິຍາມ

ຢູ່ໃນຂໍ້ນີ້ຈະສະເໜີນິຍາມ ແລະ ແນວຄິດພື້ນຖານທີ່ຈະນຳໃຊ້ໃນເຂົ້າໃນການຄົ້ນຄວ້າທົດລອງຂອງພວກເຮົາ.

ນິຍາມ 1. *TF-IDF* (Salton and Buckley 1998). ສົມມຸດ D ແມ່ນພື້ນທີ່ຂອງເອກະສານ, $d \in D$ ແລະ t ເປັນຄຳສັບໜຶ່ງທີ່ນອນຢູ່ໃນ D. *TF-IDF* ຂອງ t ຫາ d ຢູ່ໃນ D ສາມາດກຳນົດໄດ້ດັ່ງສູດຄິດໄລ່ລຸ່ມນີ້:

$$TF-IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (1)$$

ເຊິ່ງ d ເປັນເອກະສານໜຶ່ງ ແລະ t ເປັນຄາສັບໜຶ່ງຢູ່ໃນ d. $TF(t,d)$ ແມ່ນຄວາມຖີ່ຂອງຄຳສັບ t ຢູ່ໃນ d ຄືກັບການ ກຳນົດເອົາຈຳນວນຄັ້ງທີ່ t ໄດ້ປະກົດຂຶ້ນຢູ່ໃນ d. $IDF(t,D)$ ແມ່ນຄວາມຖີ່ທີ່ເກີດຂຶ້ນຈາກການປະກົດຂຶ້ນຂອງຄຳສັບ t ຢູ່ໃນເນື້ອທີ່ເອກະສານ D ຄືດັ່ງສູດລຸ່ມນີ້:

$$IDF(t,D) = \log \frac{N}{1+\left|\{d \in D : t \in d\}\right|},$$

ເຊິ່ງ N ແມ່ນຈຳນວນເອກະສານທັງໝົດຢູ່ໃນຄັງຂໍ້ມູນ, $|\{d \in D : t \in d\}|$ ແມ່ນຈຳນວນຂອງເອກະສານທີ່ມີຄຳສັບ t ປະກົດຢູ່.

ນິຍາມ 2. (Salton and Buckley 1998) ຄວາມຄ້າຍຄືກັນແບບດັ່ງເດີມ (*OrS*). ໃຫ້ $d_x = (w_{x1}, w_{x2}, \dots, w_{xT})$ ແລະ $d_y = (w_{y1}, w_{y2}, \dots, w_{yT})$ ເປັນສອງເອກະສານທີ່ຢູ່ໃນເນື້ອທີ່ຂອງເວັກເຕີ. *OrS* ລະຫວ່າງສອງເອກະສານນີ້ສາມາດຄິດໄລ່ໄດ້ດ້ວຍການໃຊ້ຟັງຊັນ *normalized cosine similarity* ດັ່ງລຸ່ມນີ້:

$$Sim(d_x,d_y)_{OrS} = \frac{\sum_{t=1}^{T}(w_{xt} * w_{yt})}{\sqrt{\sum_{t=1}^{T}(w_{xt})^2 * \sum_{t=1}^{T}(w_{yt})^2}}, \quad (3)$$

T ແມ່ນຈຳນວນຄຳສັບທັງໝົດ.

ນິຍາມ 3. ການປ່ຽນແປງສູດຄິດໄລ່ຄວາມຄ້າຍຄືກັນແບບດັ່ງເດີມ (*MoS*). ໃຫ້ $d_x = (w_{x1}, w_{x2}, \dots, w_{xT})$ ແລະ $d_y = (w_{y1}, w_{y2}, \dots, w_{yT})$ ເປັນສອງເອກະສານທີ່ຢູ່ໃນເນື້ອທີ່ຂອງເວັກເຕີ. *MoS* ລະຫວ່າງສອງເອກະສານນີ້ສາມາດຄິດໄລ່ໄດ້ດັ່ງລຸ່ມນີ້:

$$Sim(d_x,d_y)_{MoS} = \frac{\sum_{t=1}^{T}(w_{xt} - \bar{w}) * (w_{yt} - \bar{w}')}{\sqrt{\sum_{t=1}^{T}(w_{xt} - \bar{w})^2 * \sum_{t=1}^{T}(w_{yt} - \bar{w}')^2}}, \quad (4)$$

T ແມ່ນຈຳນວນທັງໝົດຂອງຄຳສັບ. $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_{xt}$ ແລະ $\bar{w}' = \frac{1}{T}\sum_{t=1}^{T} w_{yt}$.

ນິຍາມ 4. ຄວາມຄ້າຍຄືກັນດ້ານຄວາມໝາຍທີ່ໄດ້ຈາກ *word2vec* (*w2v*). ໃຫ້ $d_x = (w_{x1}, w_{x2}, \dots, w_{xT})$ ແລະ $d_y = (w_{y1}, w_{y2}, \dots, w_{yT})$ ເປັນສອງເອກະສານທີ່ຢູ່ໃນເນື້ອທີ່ຂອງເວັກເຕີ. ໂດຍອີງຕາມຫຼັກການຂອງໂມເດວ *word2vec* ແລ້ວເຮົາໄດ້ເວັກເຕີຂອງ d_x ແລະ d_y ດັ່ງລຸ່ມນີ້:

$$d_x = \frac{\sum_{t=1}^{T}(v(w_{xt}) * (TF-IDF(w_{xt}))}{\sum_{t=1}^{T}(v(w_{xt}))} \quad (5)$$

$$d_y = \frac{\sum_{t=1}^{T}(v(w_{yt}) * (TF-IDF(w_{yt}))}{\sum_{t=1}^{T}(v(w_{yt}))} \quad (6)$$

T ແມ່ນຈຳນວນທັງໝົດຂອງຄຳສັບ. $v(w_{xt})$ ແມ່ນຊຸດເວັກເຕີຂອງ w_{xt} ແລະ $v(w_{yt})$ ແມ່ນຊຸດເວັກເຕີຂອງ w_{yt} ຕາມລຳດັບ. ສະນັ້ນ ຄວາມຄ້າຍຄືກັນລະຫວ່າງສອງເອກະ ສານນີ້ ສາມາດ

ຄິດໄລ່ໄດ້ດ້ວຍການໃຊ້ *cosine similarity* ດັ່ງ ລຸ່ມນີ້:

$$Sim(d_x, d_y)_{w2v} = \frac{|d_x| \cdot |d_y|}{|d_x \cdot d_y|} \quad (7)$$

ນິຍາມ 5. ຄວາມສຳພັນດ້ານຄວາມໝາຍ (SR) (Cheng et al. 2015). ຄວາມສຳພັນດ້ານຄວາມ ໝາຍລະຫວ່າງຄຳສັບຢູ່ໃນເອກະສານແມ່ນຖືກ ເກັບກຳດ້ວຍສູດດັ່ງລຸ່ມນີ້:

$$S\ (_i \quad = \quad -\alpha \cdot \qquad\qquad +\alpha \cdot \qquad\qquad R \quad (8)$$

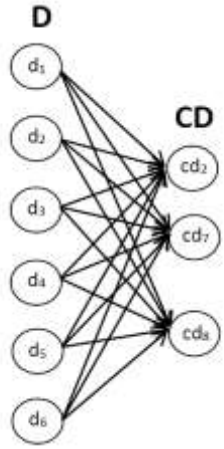
ເຊິ່ງ $PPR(t_i, t_j)$ ແມ່ນຄວາມເປັນໄປໄດ້ຂອງຄູ່ ຄວາມສຳພັນລະຫວ່າງຄຳສັບ t_i ແລະ t_j ທີ່ເກີດ ຂຶ້ນໃນເອກະສານດຽວກັນ, $PLR(t_i, t_j)$ ແມ່ນ ຄວາມສຳພັນດ້ານຄວາມໝາຍລະຫວ່າງຄຳສັບ t_i ແລະ t_j ຄືກັນກັບຄວາມເປັນໄປໄດ້ຂອງຄວາມ ສຳພັນ link-based, ແລະ α ແມ່ນການຄວບຄຸມ ລະດັບຄວາມສຳຄັນຂອງ PPR ແລະ PLR.

ນິຍາມ 6. ຄວາມຄ້າຍຄືກັນຂອງຄວາມສຳພັນດ້ານ ຄວາມໝາຍ (*SDCD*) ລະຫວ່າງເອກະສານຢູ່ໃນ ເນື້ອທີ່ເອກະສານ ແລະ ເອກະສານທີ່ຢູ່ໃນແຕ່ລະ ໝວດຂອງເອກະສານ, ສົມມຸດວ່າ D ແມ່ນເນື້ອທີ່ ເອກະສານ ເຊິ່ງບັນຈຸເອກະສານ $d_i (i = 1, 2, \ldots, N)$ ແລະ ເອກະສານທີ່ໄດ້ຈາກການ ສັງລວມເອກະສານຢູ່່ໃນໝວດຂອງເອກະສານທີ່ ໄດ້ລະບຸໄວ້ $cd_i (i = 1, 2, \ldots, M)$. ສະນັ້ນ *SDCD* ລະຫວ່າງສອງເອກະສານ d_x ແລະ d_y ສາມາດຄິດໄລ່ໄດ້ຄືດັ່ງລຸ່ມນີ້:

$$Sim(d_x, d_y)_{SDCD} = \begin{cases} 1, & if\, d(d_x, cd_i) \geq 0.5\, and\, d(d_y, cd_i) \geq 0.5 \\ 0, & if\, d(d_x, cd_i) < 0.5\, or\, d(d_y, cd_i) < 0.5 \end{cases}$$

(9) ເຊິ່ງ d_x ແລະ d_y ແມ່ນເອກະສານ, cd_i ແມ່ນໝວດເອກະສານທີ່ *i*th, $d(d_x, cd_i)$ ແມ່ນ ໄລຍະຫ່າງລະຫວ່າງເອກະສານ d_x ແລະ ໝວດ ເອກະສານທີ່ *i*th (cd_i).

ຕົວຢ່າງ: ສົມມຸດວ່າເຮົາມີເອກະສານ 6 ອັນຢູ່ໃນ ຊຸດຂໍ້ມູນທີ່ຂຶ້ນກັບໝວດເອກະສານຈຳນວນ 3 ໝ ວດ (cd_2, cd_7 ແລະ cd_8). ເຮົາຄິດໄລ່ໄລຍະຫ່າງ ລະຫວ່າງເອກະ ສານແຕ່ລະອັນກັບເອກະສານທັງ ໝົດທີ່ບັນຈຸຢູ່ໃນໝວດເອກະສານໂດຍນຳໃຊ້ຟັງ ຊັນ *cosine similarity* ດັ່ງຮູບທີ 1.



ຮູບທີ 1. ການຄິດໄລ່ ປ ແລະ ໝວດເອກະສານ (DCD)

ການສົມທຽບສອງເອກະສານໂດຍການ ນຳໃຊ້ *DCD* ແມ່ນສະແດງຢູ່ໃນຕາຕະລາງ 1.

2.2. ວິທີການ

ລາຍລະອຽດຂອງແຕ່ລະຂັ້ນຕອນແມ່ນ ໄດ້ອະທິບາຍດັ່ງລຸ່ມນີ້ ເຊິ່ງເປັນໂມເດວທີ່ ພັດທະນາຂຶ້ນມາສຳລັບການຈັດກຸ່ມເວບວິດີໂອ ໂດຍການນຳໃຊ້ຂໍ້ມູນປະເພດຂໍ້ຄວາມຢູ່ໃນຊຸດຂໍ້ ມູນ.

2.2.1. Feature extraction

ຢູ່ໃນຊຸດຂໍ້ມູນແຕ່ລະອັນຈະປະກອບດ້ວຍ 3 ພາກສ່ວນຍ່ອຍຄື: title, tag ແລະ description, ແຕ່ລະພາກສ່ວນນັ້ນບັນຈຸຄຳສັບ ຫຼື ຂໍ້ມູນທີ່ສັ້ນ ແລະ ບໍ່ມີຄວາມໝາຍຄົບຖ້ວນ. ໃນ ເວລາທີ່ພວກເຮົາແຍກຄຸນລັກສະນະຂອງຄຳສັບ (feature extraction) ພວກເຮົາໄດ້ນຳໃຊ້ເຕັກນິກ

ຫຼາຍອັນໃນຂັ້ນຕອນນີ້ ຕົວຢ່າງ: word splitter, removing word[4] ເພື່ອຕັດຄຳທີ່ບໍ່ຄວາມໝາຍອອກ ເພື່ອໃຫ້ໄດ້ຂໍ້ມູນວິດີໂອທີ່ມີຄວາມໝາຍ ແລະ ສາມາດໃຊ້ເຂົ້າໃນໂມເດວໄດ້ຢ່າງມີປະສິດທິພາບເຊັ່ນວ່າ: ຄຳຕໍ່ໜ້ານາມ, ຄາເຊື່ອມ, word stemming ແລະ tokenization.

2.2.2. Vector Space Model (VSM)

ໃນການຈັດກຸ່ມເວບວິດີໂອ, ພວກເຮົາໃຊ້ VSM ເພື່ອສົມທຽບ 2 ວິດີໂອ ດ້ວຍການໃຊ້ເວັກເຕີຂອງຄຸນລັກສະ ນະຂໍ້ມູນ. ໃນໂມເດວນີ້, ແຕ່ລະວິດີໂອແມ່ນຖືກແທນດ້ວຍເວັກເຕີທີ່ນອນຢູ່ໃນເນື້ອທີ່ຂອງເວັກເຕີລວມ ແລະ ຄິດໄລ່ຄວາມຄ້າຍຄືກັນລະຫວ່າງ 2 ວິດີໂອດ້ວຍ OrS ແລະ MoS ຕາມລຳດັບ.

2.2.3. ໝວດເອກະສານ (CD)

ຢູ່ໃນສ່ວນນີ້ຖືວ່າເປັນສ່ວນເພີ່ມເຕີມຢູ່ໃນ ໂມເດວຫຼັກຂອງພວກເຮົາ ໂດຍຄົ້ນຄິດວິທີການນຳໃຊ້ຂໍ້ມູນໃຫ້ມີຄວາມຫຼາກຫຼາຍ ແລະ ຫຼາຍດ້ານ ເພື່ອໃຫ້ເກີດຜົນປະໂຫຍດສູງສຸດ. ໃນຄັງຂອງຂໍ້ມູນ ພວກເຮົາໄດ້ຊຸ່ມເລືອກເອົາຈຳນວນຂອງເອກະສານທີ່ຈັດລຽງໄວ້ໃນແຕ່ລະໝວດ ເຊິ່ງຈານວນທີ່ຊຸ່ມມານັ້ນເທົ່າກັບຈຳນວນຂອງເອກະສານທີ່ໄດ້ເລືອກໄວ້ໃນຊຸດຂໍ້ມູນ (dataset). ໃນກໍລະນີນີ້, ເອກະສານທີ່ຊຸ່ມມາແມ່ນພວກເຮົາເອົາ ແຕ່ສ່ວນທີ່ຄາດວ່າມີຄວາມສຳຄັນຕໍ່ກັບເອກະ ສານວິດີໂອ ເຊັ່ນວ່າ: tag.

2.2.4. ຄວາມສຳພັນດ້ານຄວາມໝາຍ (SR)

ຢູ່ໃນການປະມວນຜົນທາງພາສາ (Natural Language Processing) ການວິເຄາະຄວາມສຳພັນດ້ານຄວາມໝາຍຂອງຄຳສັບແມ່ນມີຄວາມສຳຄັນທີ່ສຸດ. ພວກເຮົາໃຊ້ໂມເດວ SR ເພື່ອໃຫ້ໄດ້ຄວາມສຳພັນລະຫວ່າງຄຳສັບຢູ່ໃນ D ແລະ CD ຂອງໂມເດວ DCD (ຮູບທີ 1), ໂມເດວນີ້ ຜົນບວກທາງລິເນຍ (Linear) ຂອງຄວາມສຳພັນແບບຄູ່ ແລະ ຄວາມສຳພັນແບບ link-based ສາມາດສ້າງຄວາມສຳພັນດ້ານຄວາມໝາຍໄດ້ຢ່າງມີປະສິດທິພາບສູງ ແລະ ການຄິດໄລ່ຄວາມຄ້າຍຄືກັນລະຫວ່າງວິດີໂອແມ່ນໄດ້ນຳໃຊ້ໂມເດວ SDCD.

2.2.5. Word2vec

ເປັນເຕັກນິກໜຶ່ງທີ່ມີຮູບແບບກ່ຽວພັນກັບເຕັກນິກການໄຈ້ແຍກຄຳສັບຢູ່ໃນ word embeddings. ເຕັກນິກນີ້ໄດ້ໃຊ້ຄັງຂໍ້ມູນປະເພດຂໍ້ຄວາມທີ່ມີຂະໜາດໃຫຍ່ ເພື່ອປ້ອນເຂົ້າໃນໂມເດວເພື່ອສ້າງເນື້ອທີ່ຂອງເວັກເຕີ, ໂດຍຄາເຫຼົ່ານີ້ແມ່ນຈະບໍ່ຊ້ຳກັນຢູ່ໃນຄັງຂໍ້ມູນ ເຊິ່ງຈະຖືກກຳນົດໃຫ້ເປັນ ເວັກເຕີທີ່ສອດຄ່ອງກັບເວັກເຕີຢູ່່ໃນເນື້ອທີ່ຂອງເວັກເຕີ. ສຳ ລັບໂມເດວນີ້ ພວກເຮົາໃຊ້ໂມເດວ Skip-gram ພ້ອມກັບຊຸດເວັກເຕີມາດຕະຖານທີ່ມີຂະໜາດໃຫຍ່ Google News ໃຊ້ສຳລັບການທົດລອງ (training) ພ້ອມກັນນີ້ຂໍ້ມູນປະເພດຂໍ້ຄວາມຂອງວິດີໂອແມ່ນໄດ້ປ້ອນເຂົ້າໃນໂມເດວນີ້ເຊັ່ນດຽວກັນ.

ຕາຕະລາງ 1. ຄວາມຄ້າຍຄືກັນຢູ່ໃນໂມເດວ *DCD*.

ເອກະສານ (D)	ໝວດເອກະສານ (CD)			ຄວາມຄ້າຍຄືກັນລະຫວ່າງສອງເອກະສານ
	cd_2	cd_7	cd_8	
d_1	0.73	0.45	0.30	$Sim(d_1,d_2)_{SDCD}=1[d(d_1,cd_2)\geq 0.5 \; and \; d(d_2,cd_2)\geq 0.5]$
d_2	0.50	0.25	0.46	
d_3	0.33	0.65	0.68	$Sim(d_3,d_4)_{SDCD}=0[d(d_3,cd_8)<0.5 \; or \; d(d_4,cd_8)<0.5]$
d_4	0.29	0.38	0.43	
d_5	0.49	0.86	0.18	$Sim(d_5,d_6)_{SDCD}=1[d(d_5,cd_7)\geq 0.5 \; and \; d(d_6,cd_7)\geq 0.5]$
d_6	0.34	**0.78**	0.36	

[4] www.ranks.nl/stopwords

2.2.6. ການຮວມຄ່າຄວາມຄ້າຍຄືກັນຂອງແຕ່ລະໂມເດອ

ຜ່ານການຄົ້ນຄວ້າສັງລວມຈຸດດີ ແລະ ຂໍ້ຄົງຄ້າງຂອງວາລະສານໃນຫຼາຍໆບົດຜ່ານມາ ນັກຄົ້ນຄວ້າວິໄຈແມ່ນມີຄວາມພະຍາຍາມຄົ້ນຄິດ ແລະ ພັດທະນາໂມເດອການຈັດກຸ່່ມ ຫຼື ແບ່ງກຸ່ມ ໂດຍການໃຊ້ທັງໂມເດອດ່ຽວ ແລະ ຫຼາຍໂມເດອ ປະສົມເຂົ້າກັນເພື່ອໃຫ້ໄດ້ຜົນຮັບທີ່ດີກວ່າ, ພ້ອມນີ້ກໍ່ໄດ້ເວົ້າວ່າປະຈຸບັນນີ້ ແມ່ນບໍ່ທັນມີໂມເດອໃດ ທີ່ສາມາດຢັ້ງຢືນໄດ້ວ່າເປັນໂມເດອທີ່ໃຊ້ສຳລັບ ການຈັດກຸ່ມ ຫຼື ແບ່ງກຸ່ມໄດ້ດີທີ່ສຸດ. ມີແຕ່ຈາລອງ ໂມເດອຂຶ້ນມາໂດຍການໃຊ້ຫຼາຍໂມເດອ ທີ່ມີ ຄວາມແຕກຕ່າງກັນປະສົມເຂົ້າກັນ ເພື່ອໃຫ້ໄດ້ຜົນ ຮັບດີທີ່ແຕກຕ່າງກັນ. ເຊິ່ງແນວຄວາມຄິດຢູ່ໃນບົດ ຄົ້ນຄວ້າວິໄຈນີ້ ກໍ່ເປັນອີກວິທີໜຶ່ງ ທີ່ພະຍາຍາມ ຊອກຫາແນວທາງໃນການຈຳ ລອງໂມເດອທີ່ໃຊ້ ສຳລັບການຈັດກຸ່ມເວບວິດີໂອ ໂດຍການຄັດ ເລືອກໃຊ້ຫຼາຍໂມເດອທີ່ມີຄຸນລັກສະນະການນຳ ໃຊ້ຂໍ້ມູນໃນແຕ່ລະໂມເດອແຕກຕ່າງກັນ ເພື່ອໃຫ້ ຂໍ້ມູນໄດ້ຖືກນຳໃຊ້ຫຼາຍທີ່ສຸດ ແລະ ມີຜົນຮັບທີ່ ດີກວ່າໃນໄລຍະຜ່ານມາ. ຢູ່ໃນໂມເດອນີ້ ພວກເຮົາ ຮວມຄ່າຄວາມຄ້າຍຄືກັນຂອງແຕ່ລະໂມເດອ ເຂົ້າກັນໂດຍການຮວມແຕ່ລະຄັ້ງນັ້ນແມ່ນຂຶ້ນກັບ ຄ່າຂອງຕົວຄວບຄຸມຄວາມສຳຄັນ α ແລະ β ຂອງ ໂມເດອ ເຊິ່ງລາຍລະອຽດດັ່ງລຸ່ມນີ້:

ແທນ $Sim(d_x, d_y)_{SDCD} = S_1$, $Sim(d_x, d_y)_{w2v} = S_2$ ແລະ $Sim(d_x, d_y)_{MoS} = S_3$ ຕາມລຳດັບ ເຮົາໄດ້:

$$Sim(d_X, d_Y) = f(S_1, S_2, S_3) \quad (10)$$

$$f(S_1, S_2, S_3) = \alpha * (S_1) + \beta * (S_2) + (1 - \alpha - \beta) * (S_3) \quad (11)$$

ເຊິ່ງ $0 < \alpha, \beta < 1$ ແມ່ນຄ່າຄວາມສຳຄັນຂອງ ໂມເດອ.

2.2.7. ອາກໍຣິທຶມ

ອາກໍຣິທຶມ ສາລັບການຈັດກຸ່ມເວບວິດີ ໂອຕາມເຕັກ ນິກຂອງການຮວມກຸ່ມດ້ວຍການນາ ໃຊ້ຂໍ້ມູນປະເພດຂໍ້ຄວາມ

ເຊິ່ງສະແດງໃນອາກໍຣິທຶມ 1 ດັ່ງລຸ່ມນີ້:

ອາກໍຣິທຶມ 1.

Input: Textual information of videos.
Output: Clustering labels.
for $i \in \{title, tag, des(DS1, DS2, DS3, DS4, DS5)\}$ do
  Text Pre-processing is needed by applying TF-IDF
  scheme to find the term weights.
  Calculate the similarity matrix with OrS.
  Calculate the similarity matrix with MoS.
  for $j \in \{title, tag, des\}$ do
    Apply word2vec model.
    Calculate the semantic similarity with w2v.
  end for
  for $k \in \{title, tag, des\}$ do
    Apply SR model.
    Calculate the semantic similarity with SDCD.
  end for
  Combining similarities based on function $f(\cdot)$.
  Execute Spectral and Graphical clustering algorithms
  for getting the labels.
  Apply clustering ensemble algorithms to ensemble
  the labels.
end for

2.3. ການທົດລອງ

ພວກເຮົານຳສະເໜີຜົນຂອງການທົດລອງ ການວິໄຈໂດຍແນໃສ່ ການແກ້ໄຂບັນຫາການຈັດ ກຸ່ມ ຕາມຈຸດປະສົງ ແລະ ວິທີການຂອງການຮວມ ກຸ່ມ ໂດຍນາໃຊ້ຂໍ້ມູນທີ່ໄດ້ຈາກການອັບໂຫຼດຂອງ ຜູ້ໃຊ້ຢູ່ໃນຊຸດຂໍ້ມູນ (YouTube). ພ້ອມກັນນີ້ ເພື່ອໃຫ້ຜົນໄດ້ຮັບມີຄວາມຖືກຕ້ອງ ແລະ ມີປະ ສິດທິຜົນສູງ, ພວກເຮົາໃຊ້ວິທີການໂຮມມາຕຣິກ ຄ່າຄວາມຄ້າຍຄືກັນຂອງແຕ່ລະໂມເດອຕາມຮູບ

ແບບຂອງລິເນຍ. ຫຼັງຈາກນັ້ນ ຮວມກຸ່ມເລເບວທີ່ໄດ້ຈາກການຈັດກຸ່ມຂອງອາກໍຣິທຶມແຕ່ລະອັນ ເພື່ອໃຫ້ໄດ້ຜົນຮັບສຸດທ້າຍ. ຄວາມຈິງແລ້ວໃນໂມເດວຫຼັກຂອງພວກເຮົາປະກອບດ້ວຍ 4 ໂມເດວຍ່ອຍທີ່ແຕກຕ່າງກັນ, ແຕ່ໃນເວລາການໂຮມມາຕຣິກ ພວກເຮົາໃຊ້ພຽງແຕ່ 3 ໂມເດວເທົ່ານັ້ນ ໂດຍອີງຕາມຜົນການທົດລອງຕົວຈິງ ເຊິ່ງໄດ້ຜົນຮັບທີ່ດີກວ່າ. ສ່ວນລາຍລະອຽດຜົນຂອງການທົດລອງ ພວກເຮົາໄດ້ນຳສະເໜີໃນຂໍ້ 3.3. ສາລັບໂປຣແກຣມທີ່ໃຊ້ໃນການທົດລອງນີ້ມີຄື: C#, Python, MATLAB R2015a, ແລະ ໃຊ້ GPU ໃນການຮັນໂປຣແກຣມ.

2.3.1. ຊຸດຂໍ້ມູນ (Datasets)

ພວກເຮົາປະຕິບັດການທົດລອງ ໂດຍໃຊ້ຊຸດຂໍ້ມູນເວບວິດີໂອ (MCG-WEBV) (Cao et al. 2009) ເຊິ່ງບັນຈຸຂໍ້ມູນວິດີໂອທີ່ມີຜູ້ເຂົ້າຊົມຫຼາຍທີ່ສຸດ. ຊຸດຂໍ້ມູນນີ້ແມ່ນໄດ້ຈາກການຮວບຮວມວິດີໂອທີ່ໄດ້ຮັບຄວາມນິຍົມ ແລະ ມີຍອດວິວຫຼາຍທີ່ສຸດພາຍໃນເດືອນທັນວາ 2008 ຫາເດືອນພະຈິກ 2009 ຕາມ 15 ໝວດຂອງວິດີໂອຢູ່ໃນຢູທູບ ເຊິ່ງຂໍ້ມູນເຫຼົ່ານີ້ມີປະໂຫຍດຫຼາຍຕໍ່ກັບການນຳໄປໃຊ້ ໃນການມາຍ ນິງເວບ (web video mining) ເພາະວ່າພວກມັນບັນຈຸເນື້ອໃນກ່ຽວກັບວິດີໂອດ້ວຍຄຸນນະພາບທີ່ສູງ ແລະ ຄ່ານິຍົມຂອງຜູ້ໃຊ້. ຢູ່ໃນການທົດລອງ ພວກເຮົາໄດ້ອອກແບບການນຳໃຊ້ຂໍ້ມູນທີ່ແຕກຕ່າງກັນໃນແຕ່ຕາຕະລາງ 2. ລາຍລະອຽດ ຂອງຊຸດຂໍ້ມູນເວບວິດີໂອ ລະໂມເດວເພື່ອໃຫ້ສອດຄ່ອງກັບໂມເດວທີ່ໃຊ້, ເພື່ອໃຫ້ເກີດຜົນປະໂຫຍດສູງສຸດ ແລະ ເພື່ອເຫັນໄດ້ເຖິງຂໍ້ແຕກຕ່າງຈາກຜົນໄດ້ຮັບຕົວຈິງ ຈາກຊຸດຂໍ້ມູນດຽວກັນແຕ່ຜົນໄດ້ຮັບຕ່າງກັນ. ດັ່ງນັ້ນ, ຈຶ່ງສາມາດເຮັດໃຫ້ມີການສົມທຽບການນຳໃຊ້ຂໍ້ມູນໃນແຕ່ລະແບບ ເຊິ່ງລວມໄປເຖິງການພິຈາລະນາໃນການອອກແບບໂມເດວ ໃຫ້ມີຄວາມສອດຄ່ອງ ແລະ ໄດ້ຜົນຮັບທີ່ດີຂຶ້ນໃນການຄົ້ນຄວ້າວິໄຈໃນອະນາຄົດ. ຈາກຊຸດຂໍ້ມູນດັ່ງກ່າວນີ້, ຂໍ້ມູນແມ່ນບັນທຶກໄວ້ໃນຟາຍທີ່ຕ່າງກັນຕາມແຕ່ລະເດືອນ. ພວກເຮົາເອົາຂໍ້ມູນເຫຼົ່ານີ້ໂຮມເຂົ້າກັນເປັນຟາຍດຽວ ຫຼັງຈາກນັ້ນໃຊ້ເຕັກນິກ sampling ເພື່ອເລືອກເອົາຊຸດຂໍ້ມູນໃຊ້ສຳລັບການປະຕິການທົດລອງ ດັ່ງລາຍລະອຽດຕາຕະລາງ 2.

2.3.2. ການປະເມີນຜົນ

ເພື່ອທົດສອບປະສິດທິພາບຂອງການທົດລອງ, ພວກເຮົາໃຊ້ micro-precision (MP) (Zhou and Tang 2006) ໃນການຄິດໄລ່ຄວາມຖືກຕ້ອງຂອງການຈັດກຸ່ມໂດຍໃຫ້ຖືກຕ້ອງຕາມເລເບວຂອງພວກມັນ. MP ໄດ້ກຳນົດໄວ້ດັ່ງນີ້:

$$MP = \sum_{k=1}^{c} {}^{a_k}\!/_{n} \qquad (12)$$

ຂໍ້ມູນ	ຈຳນວນຕົວຢ່າງ	ຈຳນວນໝວດ	ໝວດ	# Features
DS1	1006	3	2, 7, 8	11700
DS2	1009	3	1, 3, 4	10112
DS3	2500	8	1, 2, 3, 4, 8, 9, 13, 15	22118
DS4	1015	3	9, 10, 14	10776
DS5	2009	6	2, 3, 4, 7, 8, 15	18324

ເຊິ່ງ n ແມ່ນຈຳນວນຂອງອັອບເจັກ ແລະ c ແມ່ນຈຳນວນຂອງກຸ່ມທີ່ຈະຈັດ, a_k ເປັນຈຳນວນອັອບเจັກທີ່ຢູ່ໃນກຸ່ມ k ເຊິ່ງກຳນົດໄວ້ຢ່າງຖືກຕ້ອງ ແລະ ສອດຄ່ອງກັບກຸ່ມ. ພ້ອມນີ້ $0 \leq MP \leq 1$ ດ້ວຍຄ່າ 1 ແມ່ນສະແດງໃຫ້ເຫັນເຖິງການຈັດກຸ່ມຂໍ້ມູນທີ່ດີທີ່ສຸດດ້ວຍການຍອມຮັບຈາກເລເບວ ຂອງກຸ່ມນັ້ນໆ.

3. ຜົນໄດ້ຮັບ

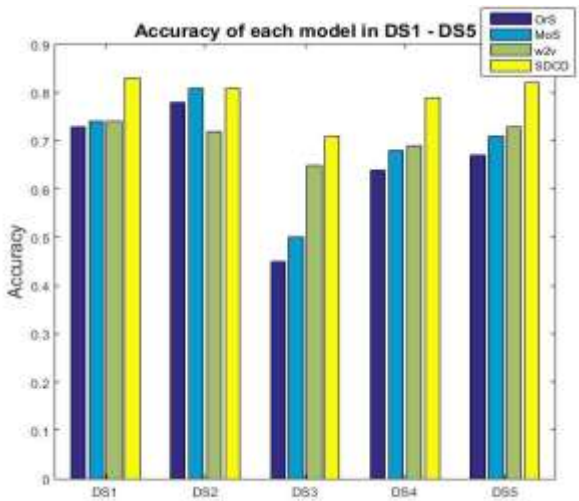
ພວກເຮົານຳສະເໜີຜົນຂອງການທົດລອງ ອອກເປັນ 2 ພາກສ່ວນຄື: ການນຳໃຊ້ໂມເດວດ່ຽວ ແລະ ໂມເດວປະສົມ (ໂຮມມາຕຣິກ). ສຳລັບໂມ ເດວປະສົມ ເພື່ອເປັນການເພີ່ມປະສິດທິພາບໃຫ້ ແກ່ໂມເດວທີ່ມີປະສິດທິພາບຕ່ຳ ພວກເຮົາໄດ້ເພີ້ມ ຄ່າຄວາມສຳຄັນໃຫ້ກັບໂມເດວນັ້ນຢູ່ໃນການໂຮມ ມາຕຣິກແບບລິເນຍ ໂດຍກຳນົດຄ່າຂອງ α ແລະ β ເທົ່າກັນ ຕົວຢ່າງ: $\alpha = 0.15$, $\beta = 0.15$ ດັ່ງນັ້ນ $(1 - \alpha - \beta) = 0.7$ ສະນັ້ນໂມເດວໃດ ທີ່ໄດ້ຮັບຄ່າ $(1 - \alpha - \beta)$ ໂມເດວນັ້ນຈະມີ ປະສິດທິພາບເພີ່ມຂຶ້ນ. ສຳລັບການທົດລອງໃນຄັ້ງ ນີ້ເຮົາໃຊ້ $\alpha = \beta = 0.1$.

3.2. ໂມເດວດ່ຽວ

ຜົນການທົດລອງຢູ່ໃນໂມເດວດ່ຽວ ເຊິ່ງ ສະແດງໃນຕາຕະລາງ 3 ແລະ ຮູບທີ 2. ຢູ່່ໃນ ຮູບທີ 2 ໄດ້ສະແດງໃຫ້ເຫັນວ່າໂມເດວ SDCD ໄດ້ຜົນໄດ້ຮັບທີ່ດີກວ່າໂມເດວອື່ນໆສຳລັບທຸກໆ ຊຸດຂໍ້ມູນ (ຍົກເວັ້ນ DS2) ໂດຍຄ່າຄວາມຖືກຕ້ອງ ໃນການຈັດກຸ່ມຂອງ DS1 ມີເຖິງ 0.83 ເຊິ່ງນັ້ນ ໝາຍຄວາມວ່າເຕັກນິກຂອງ text mining ຖ້າ ຫາກມີການພັດທະນາໂມເດວໂດຍໃຫ້ຂໍ້ມູນມີ ຄວາມສຳພັນກັນທາງດ້ານຄວາມໝາຍໃນລະດັບ ສູງ ໃນເວລານັ້ນເອງແມ່ນບໍ່ມີຄວາມຈຳເປັນຕ້ອງ ໃຊ້ແຫຼ່ງສະໜັບສະໜູນຂໍ້ມູນຈາກພາຍນອກ. ເຊິ່ງ ໂມເດວ SDCD ນັ້ນແມ່ນໄດ້ມີການວະເຄາະຂໍ້ມູນ ທີ່ໃຊ້ທັງໃນແບບການຈັບຄູ່ກັນຂອງຄຳສັບ ແລະ link-based ໂດຍໂມເດວນີ້ສາມາດຕີຄວາມໝາຍ ຂອງຄຳແຕ່ລະຄຳໄດ້ເປັນຢ່າງດີ ແລະ ເຮັດໃຫ້ການ ຈັດກຸ່ມ ຫຼື ແບ່ງກຸ່ມມີປະສິດທິພາບສູງ. ນອກຈາກ ນັ້ນ ຜົນຂອງການຈັດກຸ່ມຂອງ w2v ແລະ MoS ກໍ່ຖືວ່າໄດ້ຮັບຜົນດີເຊັ່ນດຽວກັນຕາມລຳດັບ. ດ້ວຍສາຍເຫດດັ່ງກ່າວນີ້ ຈຶ່ງເຮັດໃຫ້ມີແຮງຊຸກຍູ້ ໃນການນຳເອົາໂມເດວເຫຼົ່ານີ້ປະສົມເຂົ້າກັນ ເພື່ອ ໃຫ້ປະສິດທິພາບຂອງການຈັດກຸ່ມໄດ້ຮັບຜົນທີ່ດີກ ວ່າ.

ຕາຕະລາງ 3. ຄ່າຄວາມຖືກຕ້ອງໃນໂມເດວດ່ຽວ DS1-DS5

ໂມເດວ	OrS	MoS	w2v	SDCD
DS1	0.73	0.74	0.74	0.83
DS2	0.78	0.81	0.72	0.81
DS3	0.45	0.50	0.65	0.71
DS4	0.64	0.68	0.69	0.79
DS5	0.67	0.71	0.73	0.82
ສະເລ່ຍ (%)	*65*	*68*	*70*	*79*

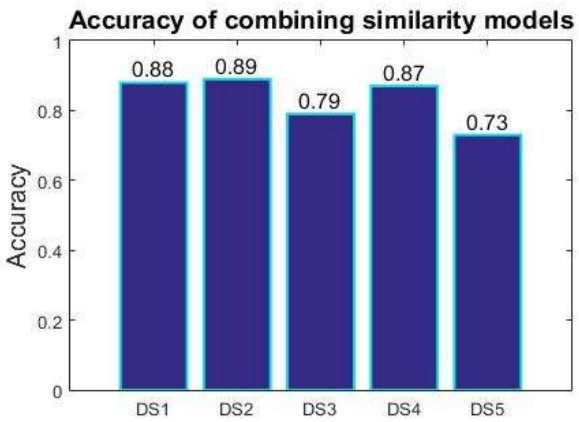


ຮູບທີ 2. ຜົນຂອງຄ່າຄວາມຖືກຕ້ອງໃນແຕ່ລະໂມເດວ

3.3. ໂມເດວປະສົມ

ໂດຍອີງໃສ່ຜົນໄດ້ຮັບຂອງການຈັດກຸ່ມຢູ່ ໃນໂມເດວປະສົມດັ່ງຮູບທີ 3 ພວກເຮົາສັງເກດ ເຫັນວ່າມີຄ່າຄວາມຖືກຕ້ອງໃນການຈັດກຸ່ມວິດີໂອ ທີ່ສູງກວ່າ ການໃຊ້ໂມເດວດ່ຽວໃນທຸກໆຊຸດຂໍ້ມູນ.

ແຕ່ວ່າມີຄ່າລຸດລົງຢູ່ໃນ DS5 ພວກເຮົາສັນນິຖານ ວ່າຂໍ້ມູນໃນ DS5 ປະກອບດ້ວຍຄຳສັບທີ່ບໍ່ມີ ຄວາມໝາຍຄົບຖ້ວນຫຼາຍເທົ່າທີ່ຄວນ ເຊິ່ງເຮັດໃຫ້ ຄວາມສຳພັນດ້ານຄວາມໝາຍບໍ່ສອດຄ່ອງກັນ, ສະນັ້ນ ຈຶ່ງເຮັດໃຫ້ຄ່າຄວາມຖືກຕ້ອງໃນການຈັດ ກຸ່ມຂອງ DS5 ລຸດລົງ. ແຕ່ເມື່ອເວລາພວກເຮົາ ສົມທຽບຄ່າສະເລ່ຍໂດຍລວມ ລະຫວ່າງໂມເດວ ປະສົມກັບໂມເດວດ່ຽວແລ້ວ ເຫັນວ່າຄ່າຄວາມ ຖືກ ຕ້ອງຂອງໂມເດວປະສົມແມ່ນ 83% ເຊິ່ງສູງ ກວ່າແຕ່ລະໂມເດວຢູ່ໃນຕາຕະລາງ 3 (MoS=68%, w2v=70% ແ ລ ະ SDCD=79%).



ຮູບທີ 3. ຜົນຂອງຄ່າຄວາມຖືກຕ້ອງໃນໂມເດວປະສົມ

3.4. ການສົມທຽບກັບໂມເດວອື່ນ

ເພື່ອເຮັດໃຫ້ຜົນໄດ້ຮັບຈາກການນຳສະເ ໜີໂມເດວຢູ່ໃນບົດວິໄຈນີ້ ມີຄວາມໜ້າເຊື່ອຖື ພວກເຮົາໄດ້ປຽບທຽບປະ ສິດທິພາບໂມເດວກັບ ເຕັກໂນໂລຊີອື່ນ ທີ່ກ່ຽວຂ້ອງກັບການນຳໃຊ້ຂໍ້ມູນ ໃນລັກສະນະດຽວກັນ ເພື່ອແກ້ໄຂບັນຫາການຈັດ ກຸ່ມ ຫຼື ແບ່ງກຸ່ມ ໂດຍສະເພາະເວບວິດີໂອ. ເຊິ່ງ ເຕັກໂນໂລຊີທີ່ໄດ້ເລືອກມາແມ່ນໄດ້ມີການອອກ ແບບໂມເດວ ສຳລັບການແບ່ງກຸ່ມຂອງເວບວິດີໂອ ດ້ວຍການນຳໃຊ້ຂໍ້ມູນທັງພາຍໃນ ແລະ ແຫຼ່ງສະ ໜັບສະໜູນຈາກພາຍນອກເຊັ່ນ: Flickr ເພື່ອຈັບ ຄູ່ກັບເນື້ອໃນທີ່ກ່ຽວຂ້ອງກັບໝວດຂອງວິດີໂອ. ຢູ່ ໃນໂມເດວການແບ່ງກຸ່ມແມ່ນໄດ້ແບ່ງເປັນ 2 ພາກສ່ວນຍ່ອຍ ຊຶ່ງນຳໃຊ້ຂໍ້ມູນປະເພດຂໍ້ຄວາມ (title, tag) ແລະ ຄຸນລັກສະນະຂອງວິດີໂອ (visual features), ຫຼັງຈາກນັ້ນໄດ້ຮວມເຂົ້າກັນ ເພື່ອເປັນຜົນໄດ້ຮັບສຸດທ້າຍ. ເຊິ່ງຄ່າສະເລ່ຍ ຄວາມຖືກຕ້ອງຂອງເຕັກໂນໂລຊີນີ້ແມ່ນປະມານ 70% ສະນັ້ນເມື່ອປຽບ ທຽບກັບປະສິດທິພາບໂມ ເດວຂອງພວກເຮົາ ເຫັນວ່າຄ່າສະ ເລ່ຍຄວາມ ຖືກຕ້ອງແມ່ນຕ່ຳກວ່າ 13% ໂດຍທີ່ຄ່າສະເລ່ຍ ຄວາມຖືກຕ້ອງສຳລັບໂມເດວທີ່ໄດ້ນຳສະເໜີຂຶ້ນ ມາແມ່ນ 83% ເຊິ່ງມີແນວໂນ້ມທີ່ດີກວ່າໂມເດວ ດັ່ງກ່າວ.

4. ວິພາກຜົນ

ຈາກຜົນຂອງການທົດລອງ, ພວກເຮົາ ໄດ້ຜົນຮັບທີ່ມີຄວາມຖືກຕ້ອງສູງພໍສົມຄວນໂດຍ ອີງໃສ່ໂມເດວປະສົມ.

(1) ແນວທາງການວິໄຈ ແມ່ນເພື່ອແນໃສ່ນຳໃຊ້ຂໍ້ ມູນປະເພດຂໍ້ຄວາມ (ຄຳສັບ) ເພື່ອຊ່ວຍໃນ ການແກ້ໄຂບັນການຈັດກຸ່ມເວບວິດີໂອ ແລະ ຊ່ວຍປັບປຸງປະສິດທິພາບຄ່າຂອງຄວາມ ຖືກຕ້ອງໃຫ້ດີຂຶ້ນ ເມື່ອທຽບກັບເຕັກໂນໂລຊີ ໃນໄລຍະຜ່ານມາ.

(2) ການກຳນົດສູດຂຶ້ນມາໃໝ່ ເພື່ອຄິດໄລ່ຄວາມ ຄ້າຍຄືກັນຂອງວິດີໂອ ຕົວຢ່າງ: MoS ແລະ SDCD ເຫັນວ່າມີຄວາມສອດຄ່ອງກັບໂມ ເດວປະສົມໄດ້ເປັນຢ່າງດີ ເຊິ່ງສະ ແດງອອກ ດ້ວຍຄ່າຂອງຄວາມຖືກຕ້ອງໃນຮູບທີ 3.

(3) ການເລືອກໃຊ້ໂມເດວ word2vec ເປັນແຫຼ່ງ ສະໜັບສະໜູນຂອງຊຸດຂໍ້ມູນ ແມ່ນເຫັນວ່າມີ ຄວາມເໝາະສົມ ແລະ ສາມາດຕອບສະໜອງ ເວັກເຕີ ທີ່ກ່ຽວຂ້ອງກັບຊຸດຂໍ້ມູນແຕ່ລະອັນ ຢ່າງມີປະສິດທິຜົນ.

(4) ຜົນໄດ້ຮັບຈາກໂມເດວປະສົມ ແມ່ນມີຄວາມ ສະຖຽນລະພາບກວ່າຜົນໄດ້ຮັບຈາກການນຳ ໃຊ້ໂມເດວດ່ຽວ ເຊິ່ງມີຄວາມຫຼາກຫຼາຍ ແລະ

ເປັນການເປີດແນວຄິດໃໝ່ໃນການວິໄຈຄັ້ງຕໍ່ໄປ.

(5) ຜົນຂອງການວິໄຈ ໄດ້ສະແດງໃຫ້ເຫັນຢ່າງຈະແจ້ງແລ້ວວ່າ ການນາສະເໜີແນວທາງນີ້ແມ່ນມີຜົນດີຕໍ່ກັບການຈັດກຸ່ມເວບວິດີໂອ ໂດຍປຽບທຽບກັບການນຳສະເໜີອື່ນທີ່ນອນຢູ່ໃນຂົງເຂດດຽວກັນ. ເຊິ່ງຂີດຄວາມສາມາດໃນການວິໄຈນີ້ ອາດຈະເປັນເຄື່ອງມືໜຶ່ງທີ່ໃຊ້ເຂົ້າໃນການແກ້ໄຂບັນຫາການສັງລວມເອກະສານຢູ່ໃນສື່ສັງຄົມອອນລາຍ ແລະ ບັນຫາທີ່ຄ້າຍຄືກັນອື່ນໆເຊັ່ນ: ການສັງລວມຂ່າວ, ການກວດຈັບສະແປມຢູ່ໃນເຄືອຂ່າຍມືເດຍ ແລະ ອື່ນໆ.

5. ສະຫຼຸບ

ຈາກການປະຕິບັດການທົດລອງ ແລະ ວິໄຈໃນຄັ້ງນີ້ສະແດງໃຫ້ເຫັນວ່າ ຂໍ້ມູນປະເພດຂໍ້ຄວາມທີ່ຜູ້ຊົມໃຊ້ສື່ສັງຄົມອອນລາຍໄດ້ອັບໂຫຼດຂຶ້ນ YouTube ພ້ອມກັບວິດີໂອນັ້ນແມ່ນສາມາດນຳໃຊ້ເຂົ້າໃນການຈັດກຸ່ມ, ແບ່ງກຸ່ມ ຫຼື ໝວດຂອງວິດີໂອໄດ້ ແລະ ຍັງຖືໄດ້ວ່າເປັນປະເພດຂໍ້ມູນທີ່ມີຄວາມສຳຄັນຕໍ່ກັບການວິໄຈໃນຂົງເຂດນີ້. ນອກຈາກນີ້ຍັງສາມາດເວົ້າໄດ້ວ່າ:

- ການກະກຽມຂໍ້ມູນ (Pre-processing) ຖືວ່າເປັນຂັ້ນຕອນສຳຄັນຂອງການວິໄຈ ທີ່ມີການນຳໃຊ້ຂໍ້ມູນປະເພດ text, ຍ້ອນວ່າວິທີການຂຽນຂອງແຕ່ລະຄົນທີ່ອັບໂຫຼດຂໍ້ມູນເຫຼົ່ານີ້ແມ່ນແຕກຕ່າງກັນ ແລະ ຂຽນໃນລັກສະນະເປັນຄຳຕິດກັນ, ໃຊ້ຕົວຫຍໍ້, ຕົວເລກ, ອີໂมจิ ແລະ ເຄື່ອງໝາຍອື່ນໆແທນຄວາມໝາຍຂອງຄຳໃດໜຶ່ງ ເຊິ່ງເຮັດໃຫ້ຫຍຸ້ງຍາກ ແລະ ບໍ່ສາມາດຕີຄວາມໝາຍຂອງຄຳສັບ ຫຼື ປະໂຫຍກໄດ້ຢ່າງຖືກຕ້ອງ ຈຶ່ງເປັນເຫດເຮັດໃຫ້ການຄິດ ໄລ່ຄ່າ ຄວາມຄ້າຍຄືກັນຂອງວິດີໂອໄດ້ຄ່າທີ່ບໍ່ຖືກຕ້ອງ ແລະ ເຮັດໃຫ້ປະສິດທິພາບໃນການຈັດກຸ່ມລຸດລົງ. ດັ່ງ ນັ້ນ, ການນຳໃຊ້ເຕັກນິກຕ່າງໆເຂົ້າໃນຂັ້ນຕອນນີ້ແມ່ນເພື່ອໃຫ້ໄດ້ຂໍ້ມູນທີ່ມີເນື້ອໃນຈະແจ້ງ,ມີຄວາມໝາຍແລະ ສາມາດເປັນປະໂຫຍດຕໍ່ກັບການນຳໄປໃຊ້ ໃນການຈັດກຸ່ມ.
- ການພິຈາລະນານຳໃຊ້ໂມເດວທີ່ມີຄວາມເໝາະສົມ ແລະ ສອດຄ່ອງກັບຂໍ້ມູນ ແມ່ນມີຄວາມສາມາດໃນການຈຳ ແນກ ແລະ ການພັນລະນາຄວາມໝາຍຂອງຄວາມສຳພັນລະຫວ່າງຄຳສັບ ເຊິ່ງຖືເປັນປັດໃຈສຳຄັນໃນການຈັດກຸ່ມຂໍ້ມູນເວບວິດີໂອ.
- ການນຳສະເໜີໂມເດວທີ່ມີການປະສົມກັນຢ່າງກົມກຽວ ແລະ ຖືກຕ້ອງນັ້ນ ເປັນເຄື່ອງມືທີ່ມີຄວາມສາມາດໃນການຈັດກຸ່ມໄດ້ຢ່າງມີປະສິດທິພາບ ແລະ ມີຄ່າຄວາມຖືກຕ້ອງສູງ. ຢູ່ໃນໂມເດວປະສົມ ພວກເຮົາພຽງແຕ່ນຳໃຊ້ 3 ໂມເດວຍ່ອຍເທົ່ານັ້ນ, ແຕ່ຖ້າພວກເຮົາພິຈາລະນາໂມເດວ OrS ເຂົ້າມາໂດຍສະລັບປະສົມກັບໂມເດວອື່ນໆ ແລະ ພ້ອມທັງກຳນົດຄ່າຂອງ α ແລະ β ໃຫ້ມີຄວາມຫຼາກຫຼາຍ ແລະ ທາການວິເຄາະຜົນໄດ້ຮັບໃນແຕ່ລະຄ່ານັ້ນເຮົາອາດຈະເຫັນຜົນທີ່ມີຄວາມແຕກຕ່າງກັນອອກໄປ ແລະ ສາມາດໄດ້ຄ່າທີ່ດີທີ່ສຸດຂອງ α ແລະ β.

ເຖິງແມ່ນວ່າ ຜົນຂອງການທົດລອງຈະໄດ້ດີແລ້ວກໍ່ຕາມ ແຕ່ຂໍ້ມູນຍັງບໍ່ມີຄວາມສະຖຽນລະພາບຢ່າງພຽງພໍ ຍ້ອນການອັບໂຫຼດຂໍ້ມູນບໍ່ໄດ້ກຳນົດລະດັບຂອງຜູ້ຊົມໃຊ້ ຈຶ່ງເຮັດໃຫ້ຂໍ້ມູນບໍ່ສົມບູນ. ສະນັ້ນ, ມັນຈາເປັນຕ້ອງໄດ້ຮັບການສະໜັບສະໜູນຈາກແຫຼ່ງຂໍ້ມູນພາຍນອກ. ໃນແຜນການວິໄຈໃນຕໍ່ໜ້າ ພວກເຮົາຈະພິຈາລະນານຳໃຊ້ຂໍ້ມູນເພີ່ມເຕີມທີ່ມາພ້ອມກັບຊຸດຂໍ້ມູນນີ້ ແລະ ຄົ້ນຫາເຕັກນິກການໝູນໃຊ້ຂໍ້ມູນທີ່ມາຈາກສື່ສັງຄົມອອນລາຍເປັນຕົ້ນແມ່ນ: ImageNet, Wikipedia, WordNet ແລະ ອື່ນໆ ໂດຍສົມທົບການພັດທະນາໂມເດວການຈັດກຸ່ມໃໝ່.

6. ຄຳຂອບໃຈ

ຂ້າພະເຈົ້າ ຂໍສະແດງຄວາມຮູ້ບຸນຄຸນມາ ຍັງຄະນະນຳ ຄະນະວິສະວະກຳສາດ, ຫ້ອງການ ຄົ້ນຄວ້າ ແລະ ບໍລິການວິຊາການ ແລະ ມະຫາວິທະຍາໄລສຸພານຸວົງ, ຂໍຂອບໃຈມາຍັງ Prof. Tianrui Li. PhD ທີ່ໃຫ້ຄຳປຶກສາ ແລະ ຄວາມຄິດເຫັນຕ່າງໆທີ່ປະກອບຢູ່ໃນບົດວິໄຈນີ້ ພ້ອມທັງອຳນວຍຄວາມສະດວກໃນການນຳໃຊ້ ຫ້ອງແລັບ ເພື່ອທາການທົດລອງໃນຄັ້ງນີ້ຈົນ ສຳເລັດດ້ວຍດີ.

7. ເອກະສານອ້າງອີງ

Cao, J., Zhang, Y. D., Song, Y. C., Chen, Z. N., Zhang, X., & Li, J. T. (2009). MCG-WEBV: A benchmark dataset for web video analysis. *Beijing: Institute of Computing Technology*, *10*, 324-334.

Cheng, X., Miao, D., & Wang, C. (2015). A link-based approach to semantic relation analysis. *Neurocomputing*, *154*, 127-138.

Chen, Z., Cao, J., Song, Y., Zhang, Y., & Li, J. (2010). Web video categorization based on Wikipedia categories and content-duplicated open resources. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1107-1110). ACM.

Cilibrasi, R., & Vitanyi, P. (2006). Automatic meaning discovery using Google. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Cilibrasi, R. L., & Vitanyi, P. M. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, *19*(3).

Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 835-850.

Huang, T. C. K., Chen, Y. L., & Chen, M. C. (2016). A novel recommendation model with Google similarity. *Decision Support Systems*, *89*, pp. 17-27.

Jiang, Y. G., Ngo, C. W., & Yang, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval* (pp. 494-501). ACM.

Ju, R., Zhou, P., Li, C. H., & Liu, L. (2015). An efficient method for document categorization based on word2vec and latent semantic analysis. *IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM)*, (pp. 2276-2283).

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. *International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th* (pp. 136-140).

Mahmood, A., Li, T., Yang, Y., Wang, H., & Afzal, M. (2015). Semi-supervised evolutionary ensembles for Web video categorization. *Knowledge-Based Systems*, *76*, 53-66.

Mekthanavanh, V., & Li, T. (2016). Social Web Videos Clustering Based on Ensemble Technique. In *International Joint Conference*

on Rough Sets (pp. 449-458). Springer, Cham.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Song, Y., Zhao, M., Yagnik, J., & Wu, X. (2010). Taxonomic classification for web-based videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 871-878.

Schindler, G., Zitnick, L., & Brown, M. (2008). Internet video category recognition. In *CVPRW'08: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008*. University of Bath.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles acknowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*(Dec), 583-617.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513-523.

Toderici, G., Aradhye, H., Pasca, M., Sbaiz, L., & Yagnik, J. (2010). Finding meaning on youtube: Tag recommendation and category discovery. In *IEEE Congerence on Computer Vision and Pattern Recognition (CVPR)*, pp. 3447-3454.

Topchy, A., Jain, A. K., & Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12), 1866-1881.

Wang, Z., Zhao, M., Song, Y., Kumar, S., & Li, B. (2010). Youtubecat: Learning to categorize wild web videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 879-886.

Wu, X., Zhao, W. L., & Ngo, C. W. (2009). Towards google challenge: combining contextual and social information for web video categorization. In *Proceedings of the 17th ACM international conference on Multimedia* (pp. 1109-1110). ACM.

Wu, X., Ngo, C. W., Zhu, Y. M., & Peng, Q. (2012). Boosting web video categorization with contextual information from social web. *World Wide Web*, *15*(2), 197-212.

Yang, L., Liu, J., Yang, X., & Hua, X. S. (2007). Multi-modality web video categorization. In *Proceedings of the international workshop on Workshop on Multimedia Information Retrieval* (pp. 265-274). ACM.

Yuan, X., Lai, W., Mei, T., Hua, X. S., Wu, X. Q., & Li, S. (2006). Automatic video genre categorization using hierarchical SVM. In *IEEE International Conference on Image Processing*, pp. 2905-2908.

Zhou, Z. H., & Tang, W. (2006). Clusterer ensemble. *Knowledge-Based Systems*, *19*(1), 77-83.